Maximizing the Spread of Influence in Temporal Social Networks

Wrocław University of Technology

Radosław Michalski

Department of Computer Science and Management
Institute of Informatics

Wrocław University of Technology

A thesis submitted for the degree of

Doctor of Philosophy

Wrocław, 2014

Abstract

People do not live in isolation. They are constantly exchanging information, they form groups, establish families. Nowadays they are also overwhelmed by the constant media stream. All of the above influence ones opinions, visions and attitudes. By looking at human relationships from a social network perspective, these opinions spread through the network convincing some people to share them and pass to their neighbours, and others to discard or disagree with them. This process, named "the spread of influence", is happening all the time and following it is one of the challenges. The other challenge, definitely more difficult, is to predict whether a particular opinion will spread across the network. The third aspect discussed in this dissertation, is how to choose an initial set of nodes - the seed - in order to maximise the overall spread of influence in the social network.

In contrast to the state-of-the-art in this area, the dissertation concerns a problem of the dynamics in social networks. As we amplify or reduce the intensity of contacts with others over time or sometimes even completely change our environment, the assumption that the social network mapping these relationships is static is too strong simplification of reality. That is why the temporal aspects of networks are influencing the whole process strongly and discarding that information may lead to wrong conclusions about whom to influence in the beginning. Moreover, as contacts between people are ordered in time, it is impossible to look at this process as at a reversible one - information cannot travel against time. The static approach assumes this as well by saying that relationships established some time ago, which ceased over time, may be used for passing information to others. This assumption is simply wrong.

In this work it has been studied how the spread of influence behaves in temporal social networks and whether it is possible to introduce a new approach outperforming typical strategies in this area by making use of the temporal nature of networks. The model of social influence studied in this dissertation is Linear Threshold, one of two most commonly used in the research on social influence.

In order to face common problems in the area, a new method providing better results is proposed and evaluated in this dissertation - the *tInf* method. It proves that instead of simplifying the reality it is better to make use of it in order to achieve more satisfying results. However, it should be stated, that this approach is not a win-win solution, since it also introduces a new set of challenges that should be solved to take full advantage of it.

Streszczenie

Ludzie nie żyją w odosobnieniu. Nieustannie wymieniają się informacjami, tworzą grupy, zakładają rodziny, a także w coraz większym stopniu przytłaczani są informacjami płynącymi z mediów. Całe to otoczenie wpływa na ludzkie opinie, zdanie i nastawienie. Obserwując nasze otoczenie z perspektywy sieci społecznych, te opinie rozprzestrzeniają się w sieci przekonując niektórych do ich przyjęcia i przekazania dalej a innych do odrzucenia lub braku akceptacji. Ten proces, nazwany rozprzestrzenianiem się wpływu dzieje się nieustannie i jego śledzenie jest tylko jednym z wyzwań. Kolejne, zdecydowanie bardziej złożone, to próba określenia czy konkretna opinia rozprzestrzeni się w sieci. Następny problem, analizowany w tej rozprawie, dotyczy wyboru początkowego zbioru węzłów (aktorów) w sieci społecznej mających maksymalizować wpływ w tej sieci.

W przeciwieństwie do stanu wiedzy w tej dziedzinie, w rozprawie podjęto temat dynamiki sieci społecznych. Ponieważ ludzie intensyfikują lub zmniejszają kontakty z innymi w czasie lub czasem całkowicie zmieniają swoje otoczenie, założenie, że sieć społeczna odwzorowująca te relacje jest statyczna, jest zdecydowanie zbyt dużym uproszczeniem rzeczywistości. Z tego właśnie powodu czasowe własności sieci w dużym stopniu wpływają na cały proces a ignorowanie tego faktu może doprowadzić do złych wniosków w zakresie wyboru początkowych węzłów. Ponadto, ponieważ kontakty międzyludzkie występują w pewnej kolejności, niemożliwe jest spojrzenie na ten proces jako na odwracalny. Podejście przyjmowane w sieciach statycznych zakłada, że do rozprzestrzniania wpływu można użyć kontaktów zawartych dawno temu i już nieużywanych - jest to podejście zdecydowanie błędne.

W pracy rozważano w jaki sposób rozprzestrzenianie się wpływu zachodzi w dynamicznych sieciach społecznych i czy możliwe jest zaproponowanie nowego podejścia, które sprawdzi się lepiej od dotychczasowych algorytmów poprzez wykorzystanie czasowych aspektów sieci społecznych. Model rozprzestrzeniania się wpływu w sieciach analizowany w pracy to *Linear Threshold*, jeden z dwóch najczęściej wykorzystywanych w badaniach z tego zakresu.

W celu rozwiązania problemów występujących w studiowanym obszarze, w niniejszej rozprawie zaproponowano nowe podejśie - metodę *tInf* - zapewniającą lepsze rezultaty oraz przeprowadzono jej analizę. Wykonane badania dowodzą, że zamiast upraszczać rzeczywistość lepiej jest wykorzystać zmienność sieci w celu uzyskania lepszych wyników. Niemniej jednak należy nadmienić, że zaproponowane podejście nie jest panaceum na wszystko, gdyż także wprowadza pewne komplikacje, głównie w zakresie ilości danych niezbędnych do przetworzenia. Aby więc w pełni wykorzystać korzyści płynące z nowej metody, należy przezwyciężyć i te problemy.

Contents

List of Figures

List of Tables

List of Algorithms

Nomenclature

Symbols

$\Phi(0)$ Set of initially influenced nodes - seed

$\Phi(T_p)$ Set of influenced nodes in time T_p

θ_v Threshold level for node v

D_z Dataset z

K Number of windows in the training temporal social network

K' Number of windows in the testing temporal social network

m A propagation model of influence

N_v^{inf} Set of influenced neighbours of node v

$N_{i_{T_p}}$ Neighbouring nodes of node v_i in time window T_p

P^m A set of parameters of a propagation model m

SI or $SI^{SN}(SN, m, P^m, SC, \Phi(0))$ Spread of Influence for social networks, i.e. the number of finally influenced nodes

$SI^{TSN^K}(TSN^K, m, P^m, p, \Phi(0))$ Propagation model for TSN

TES_{T_p} Time-limited Event Sequence

TSN^K Temporal Social Network consisting of K Social Networks

SC Stop Condition

SN Social Network

ES Event Sequence

Selected Acronyms

DNA Dynamic Network Analysis

IC Independent Cascade Model

LT Linear Threshold model

NG Naming Game

SNA Social Network Analysis

VM Voter Model

Chapter 1

Introduction

1.1 General Characteristic of the Research Domain

This dissertation focuses on the problem of the spread of influence in social networks. The spread of influence is a process of influencing individuals with some attitude, idea or opinion that is initiated by an external entity or entities, namely influencers. They have different powers of influence, since some people are more influential, and some less. Moreover, individuals have their own threshold of becoming influenced, i.e. individuals differ in resistance to pressure from others.

Unlike all the previous works in this area, this thesis considers the problem of dynamic networks, i.e. networks that change over time. Those are the real networks we are part of, since our activity is not constant. We meet new people, change the frequency of contacts, abandon some relationships etc. It is not always the case that networks of this type are static and there is no argumentation that they should be treated like that. What is more, current research shows that by using the simplification of time-aggregated networks, the outcomes of the process are simply misleading.

This thesis aims to provide a full introduction to the area of the problem of spread of influence in dynamic or, more often referred to as temporal, social networks. It focuses on the problem of maximizing the spread of influence in the temporal setup, attempting to answer the question of whether it is possible

to benefit from the temporal information about nodes' past activity in order to maximize the spread, i.e. the number of infected nodes after certain condition is reached, e.g. after some time. The model of influence considered in this thesis is Linear Threshold which is one of the most popular models used in the research.

1.2 Applications of Maximizing the Spread of Influence in Social Networks

The idea of social influence is a general one, i.e. it may have either positive or negative applications. As the problem considered in this thesis focuses on maximizing the spread of influence, the outcome is the highest number of influenced individuals after a certain period of time. This is why this method can be applied in variety of domains - some exemplary applications are presented below.

One of the most popular application area is marketing. Since the era of undirected or multicast marketing is changing towards more directed one, new marketing techniques are trying to make use of so-called network value of individuals. The network value is the potential of a person to recommend the product to its friends or relatives in case if this person becomes influenced. So, in order to maximize the revenue with a given budget some individuals are targeted and become seeds, i.e. the marketing campaign will target them first. Targeting may be of different nature, such as sending trials of products, gifts, personalized ads etc., but the expected outcome remains the same - maximization of the spread and, in the end, the revenue from this campaign. Nowadays research is showing that the campaigns of this type become more and more popular, since viral marketing or gossip marketing are the current buzzwords.

Another area may be the social one, e.g. promoting good habits or conducting social campaigns. In order to maximize the engagement of individuals, targeting the society may be conducted to choose the most influential nodes. An interesting case may be for instance the promotion of vaccinating children or washing hands before eating.

Naturally, there exists another important application of this idea, politics. Effective influencing individuals to vote on particular parties or candidates is an

essential goal for politicians and it may be accomplished by the methods presented in this thesis.

However, the list is not limited only to positive or neutral applications, since, typically, this is a double-edged sword. In order to start a riot against governments, or a rebellion, one may use the presented techniques to target and influence the individuals that convince others. On the other hand, it may be used to neutralize some individuals in the warfare time.

By saying that, it is concluded that the method presented in this thesis does not focus on a particular application. This is a general technique of maximizing the spread of influence in temporal networks without sticking to a particular application. Naturally, the author of this dissertation hopes that the applications of it will only be beneficial both to the society and science.

1.3 Dissertation's Motivation, Aim and Contribution

The most important motivation for this work is that since the beginning of solving the problem of maximizing the spread of influence, only static networks have been considered. Researchers have proposed many techniques that try to accomplish this task in the static environment. However, the assumption that modelling this dynamic process in static networks is beneficial, has come to be weaker and weaker in last few years. A number of works summarized in Holme and Saramäki [2012] clearly show that it does not hold any more, and there is no argument to stick to the static (time-aggregated) social networks when modelling dynamic processes. In order to overcome this limitation, this thesis shifts the problem to the dynamic environment and attempts to evaluate how this dynamics influences the outcomes compared to the static approach. So, right now, there are two dynamic processes embedded: the dynamics of networks and the dynamics of the spread of influence. Neither of them is considered a trivial one to analyse separately, and a combination of them makes the challenge even more exciting.

This dissertation aims to present a new and scalable solution for maximizing the spread of influence for temporal social networks. In order to do this it is

required to prepare a proper background, since the domain of temporal social networks is rather new. Firstly, the necessary definitions are introduced: Event Sequence, Social Network and Temporal Social Networks are defined in Chapter 2 in order to have a common understanding of the main concepts. Next, Chapter 3 presents three important processes that happen in social networks related to the spread: diffusion of information, diffusion of innovations and social influence. Understanding the differences between these is important in order to process to the idea of maximizing the spread of influence, since other factors are involved in each of these processes. Naturally, as the thesis title suggests, further work is devoted to the process of social influence. This process is described and studied in terms of models in the same chapter. Chapter 4 presents the state-of-the-art and current challenges in the area of social influence maximization, both for static and dynamic networks. Chapter 5 is the first chapter that introduces the experimental set-up. It provides the information on how the experiments will be conducted and presents initial properties of evaluated datasets, especially related to their dynamics. These properties were further exploited by the *tInf* method introduced in this thesis that maximizes the spread of influence in temporal networks. This method along with the experiments is introduced in Chapter 6 and it proves that instead of ignoring temporal information in social networks, this temporal information may be successfully exploited.

Trying to summarize the contributions of this thesis, they are enumerated below, ordered by their importance - from the most important to the least important. An extensive summary of contributions of this dissertation is presented in the last chapter.

1. *The novel model of temporal spread*
   In this dissertation a new way of showing how to deal with the spread of influence in temporal networks was presented. The proposed model splits the data into training and testing periods and uses the temporal social network in order to evaluate the seeding strategies. This is why it provides an universal framework for the whole process.

2. *The tInf method of maximizing the spread of influence in temporal networks*
   The *tInf* method introduced in this dissertation takes the advantage of the

nodes' temporal activity in order to build a seed set to maximize the spread of influence.

3. *Experimental verification of the tInf method*
   The *tInf* method has been experimentally evaluated on five real world datasets. The results reveal the advantage of the proposed approach.

4. *The development of the software framework for spread of influence in temporal social networks*
   A dedicated software framework has been developed. It allows to perform multiple tasks related to the area of spread of influence and temporal social networks. This framework is suitable to perform the experiments in the cluster environment in order to reduce the run time and perform more extensive research.

5. *Ordering the domain*
   This dissertation also attempts to introduce common definitions for new and differently used terms related to the area of temporal social networks and spread of influence.

Summarizing, this thesis may be considered as, more or less, complete introduction to the problem of maximizing the spread of influence in temporal social networks by providing the definitions of terms, stating the challenge, introducing the method and evaluating it.

In order to begin the challenge of maximizing the spread of influence in temporal social networks, the following chapter introduces the above mentioned definitions.

Chapter 2

Social Networks and Temporal Social Networks

2.1 Introduction

The role of this chapter is to present the concepts which play the role of underlying layers for dynamic processes studied in Chapter 3. These dynamic processes are diffusion of information, diffusion of innovations and social influence that occur in the network environment. This chapter introduces a number of definitions which serve as a background for these processes. Starting with an idea of event sequence representing contacts or collaborative, social activities between individuals, a definition of a social network is provided. Afterwards, the concept of a temporal social network is defined. It is a time-respecting representation of human interactions in which dynamic processes, such as spread of influence, may be modelled. It makes the whole configuration as close as possible to real-life scenarios.

2.2 Event Sequence

Our typical everyday activities include meeting people, talking to them, exchanging emails, making phone calls, using social media platforms and instant messengers to stay in touch with our friends, relatives and colleagues. Each of those

activities often leaves a trace that may be used for analysing multiple aspects of communication. In this thesis, those interactions between individuals are named Event Sequence (ES) and it is defined as follows:

Definition 1. *An Event Sequence is a tuple* $ES = (V^e, EV)$*, where* $V^e = \{v_1^e, \ldots, v_{n^e}^e\}, n^e \in \mathbb{N}_+$ *is the set of social entities and* $EV = \{ev_1, \ldots, ev_{k^{ev}}\}, k^{ev} \in \mathbb{N}_+$ *is the finite set of events (contacts) between them. Each event* ev_{ijkl} *is a tuple:* $ev_{ijkl} = (v_i^e, v_j^e, t_k^e, id_l^e)$*, where* $v_i^e, v_j^e \in V^e$*,* $v_i^e \neq v_j^e$ *and* $t_k^e \in T^e$*. Here,* T^e *represents a discrete time dimension consisting of timestamps* $T^e = \{t_1, \ldots, t_m^t\}$*,* $m^t \in \mathbb{N}_+$ *in which a particular event occurred or is assigned to. The set* $ID = \{id_1, \ldots, id_{n^{id}}\}$ *contains unique event identifiers* $id_{e^{id}} \in ID, n^{id} \in \mathbb{N}_+$*. The set of nodes* V^e *cannot possess any isolated nodes, i.e.* $\underset{i}{\forall}(v_i^e \in V^e \Leftrightarrow \underset{jkl}{\exists}(e_{ijkl} \in EV \vee e_{jikl} \in EV))$*.*

This definition allows to formalize events binding individuals and it introduces the most granular type of information about those events. It means that $ev_{ijkl} = (v_i^e, v_j^e, t_k^e, id_l^e)$ may be interpreted as the fact of sending an email with the identifier id_l^e by an individual v_i^e to v_i^e in time t_k^e. It should be remembered that the content or type of events is not considered here, only the fact of their occurrence. An exemplary ES could be the email server log list or the phone calls registry. Naturally, such an event sequence may be also obtained by conducting surveys and asking respondents to provide the information about who they met and talked to.

In the above definition ID is used for labelling events which were not of unicast type, i.e. when one individual v_i^e contacts towards multiple ones simultaneously, for instance by sending a single email with id_l^e to many recipients. If so, many events ev_{ijkl} are created in ES with the same initiator v_i^e, same timestamp t_k^e and same identifier id_l^e separately for each distinct recipient v_j^e. In order to bind those events and know that they had the same origin, the identifier id_l^e is needed. As sometimes, it is crucial to measure the importance of relations, such a quantitative may be lower if the event was a part of multicast communication and higher if it was a unicast (Kazienko et al. [2009]).

Please note that the set V^e consists only of nodes that have to be a part of any event, being either its initiators or recipients. The restriction $v_i^e \neq v_j^e$ prevents

from loops. Events are strictly directed, i.e. if ev_{ijkl} exists, it does not mean that ev_{jikl} also exists.

2.3 Social Network

The event sequence introduced above may be considered as a raw data. By using this as the basis for reasoning, it is hard to perform more sophisticated aggregated analysis. In order to make it more adequate for deeper analysis, it is transformed into the Social Network (SN).

2.3.1 Definition

Definition 2. *A Social Network SN on Event Sequence $ES = (V^e, EV)$ is a tuple $SN = (V, E)$, where $V = \{v_1, \ldots, v_n\}, n \in \mathbb{N}_+$ is the set of vertices and $E = \{e_1, \ldots, e_{k^e}\}, k^e \in \mathbb{N}_+$ is the set of edges between them. Each vertex $v_i \in V$ represents an individual v_i^e from Event Sequence and each edge e_{ij} corresponds to the directed social relationship from v_i to v_j, such that $E = \{(v_i, v_j, w_{ij}) : v_i \in V, v_j \in V, v_i = v_i^e, v_j = v_j^e$ and $\underset{ij}{\forall}(\underset{kl}{\exists} ev_{ijkl} \in EV \Leftrightarrow e_{ij} \in E), w_{ij} \in [0, 1]\}$. Here, value $w_{ij} = \frac{n_{ij}^e}{n_i^e}$ denotes the importance (weight, strength) of the relationship between individuals, such that n_{ij}^e is the number of events ev_{ijkl} from v_i^e to v_j^e in ES (regardless k, l) and n_i^e is the number of all events initiated by v_i^e (outgoing from).*

The social network introduced above is defined on the event sequence, i.e. all the nodes that appear in SN have to belong to V^e, actually $V = V^e$, and all the relationships represented by edges E need to be derived from this event sequence. The social network, in fact, aggregates the event sequence file into a directed and weighted graph. Moreover, the weight w_{ij} represents the importance of relationship between v_i and v_j expressed as a fraction of a number of events from v_i to v_j divided by the number of events initiated by v_i.

Due to the fact that the Social Network SN was defined on Event Sequence ES and is based on Definition 1, it is impossible to have isolated nodes in V as well as loops. Moreover, the social network is also directed, i.e. if an edge e_{ij} exists, it does not imply the existence of e_{ji}.

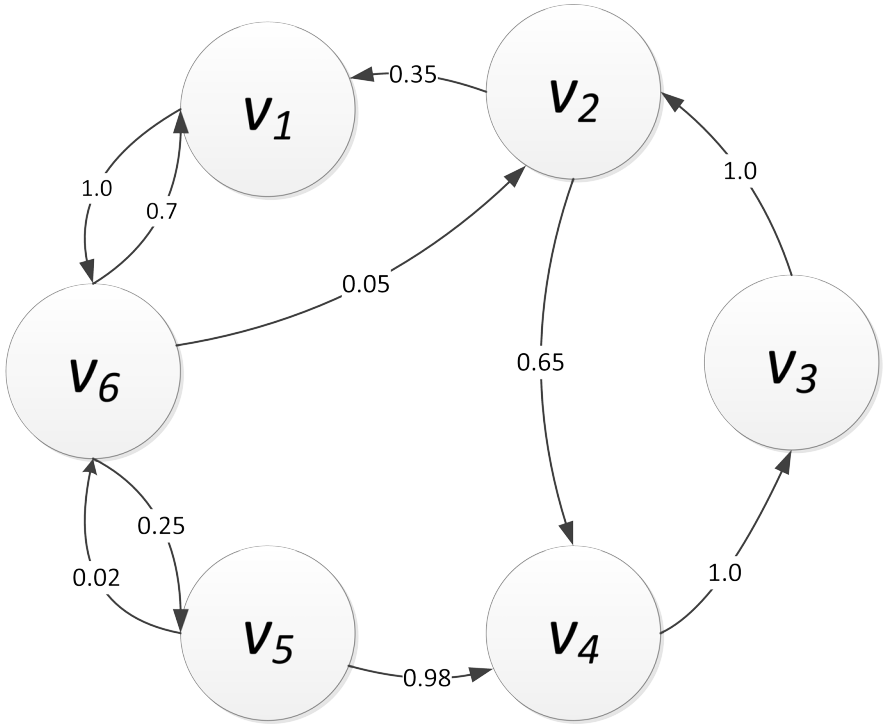


Figure 2.1: An exemplary social network. It consists of six vertices representing individuals and edges reflecting relations. The network is weighted and directed.

This definition enables to perform complex analyses that are mostly derived from the graph theory (Prell [2011]). In the above definition, it is assumed that the relationship is directed, i.e. from v_i to v_j, but there are also variants where the direction is not considered (undirected social networks) or the importance of relation is omitted (unweighted social networks), see Wasserman and Faust [1994]. In this dissertation, further studies are based on the above definition of the weighted and directed social network. An exemplary social network that is compliant with Definition 2 is presented in Figure 2.1.

2.3.2 Limitations

As it may be concluded from Definition 2 of SN, the time dimension is not used here, so it is hard to say whether some edges correspond to old events or recent ones. This is because the timestamp of the event was not considered when transforming the event sequence into the social network. This kind of social network

is often called *static* or *time-aggregated* social network (Butts [2009]). In order to overcome this limitation, a number of strategies are proposed. Firstly, instead of aggregating contacts between the same pair of individuals in one edge, multiple timestamped edges are allowed (*contact sequence*, see Holme and Saramäki [2012]). This representation is a straightforward transformation from event sequence to the social network, but it requires the introduction of completely new methods for computing social network measures. The typical ones could not deal with timestamped edges.

Another approach is to create a sequence of static social networks which is time-ordered and every such social network aggregates contacts from a given timeframe (Bródka et al. [2013]). This approach is called a Temporal Social Network (TSN). It could be considered as a trade-off between contact sequence and time-aggregated networks, since it allows the use of standard and well established structural measures while computing the importance of nodes. The temporal aspects are also preserved, because of the time order that is used in a sequence. Before defining TSN, it is required to define a Time-limited Event Sequence (TES) which will simplify the definition of TSN.

2.4 Time-limited Event Sequence

Definition 3. *A Time-limited Event Sequence* $TES_{T_p} = (V^e_{T_p}, EV_{T_p})$ *on Event Sequence ES is the Event Sequence limited only to events within a given period* $T_p = (t_p, t'_p, closure_l, closure_u)$*, i.e.* $EV_{T_p} = \{(v^e_i, v^e_j, t^e_k, id^e_l) : t^e_k \in T_p\}$*. The* $closure_l$ *refers to the lower bound of the period and can be either of type "[" or "(". Upper bound is described by* $closure_u$ *and also can be either "]" or ")".*

Please note that this time restriction filters the events to the certain period T_p, but it also refers to nodes, i.e. nodes that are not involved in any activity (event) within the given period (isolated ones) are removed. Obviously, it is possible that if there are no events in ES for the given period T_p, TES will contain empty sets of individuals and events.

Moreover, at this point it is also important to underline one more property of TES. If the period T_p covers all events in ES, the set $V^e_{T_p}$ will be exactly the

same as the set V^e. When the T_p is shorter, the set of individuals $V^e_{T_p}$ may not necessarily be the same. This will be the case if the period covered by T_p does not include the individuals that were initiators or recipients only in events outside T_p. In such case $V^e_{T_p} \subset V^e$, but generally $V^e_{T_p} \subseteq V^e$. The same remark applies to EV_{T_p}.

Period $T_p = (t_p, t'_p, closure_l, closure_u)$ may represent four types of ranges: $[t_p; t'_p]$, $[t_p; t'_p)$, $(t_p; t'_p)$, $(t_p; t'_p]$, depending on the closure types. It means that $t^e_k \in [t_p; t'_p]$, $t^e_k \in [t_p; t'_p)$, $t^e_k \in (t_p; t'_p)$, $t^e_k \in (t_p; t'_p]$, respectively.

2.5 Temporal Social Network

Definition 4. *A Temporal Social Network TSN^K on Event Sequence ES is a sequence of time-ordered component Social Networks SN_p, such that $TSN^K = (SN_1, \ldots, SN_p, \ldots, SN_K)$, $K \in \mathbb{N}_+$. A component Social Network SN_p is extracted from Time-limited Event Sequence TES_{T_p} (see Definition 3). The time order is non-descending, i.e. $\underset{1 \leqslant p < K}{\forall} t_p \leqslant t_{p+1}$ and $\underset{1 \leqslant p < K}{\forall} t'_p \leqslant t'_{p+1}$.*

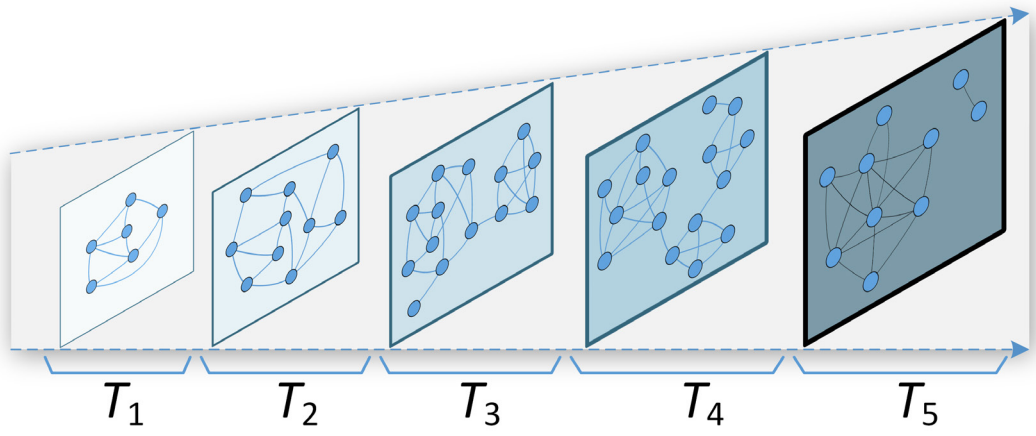


Figure 2.2: An illustration of the temporal social network as defined in this thesis.

Due to the fact that multiple events (common social activities) might occur between v_i and v_j within a single time window T_p, the relationship is obtained by aggregation of these events. Therefore, each SN_p can be treated as a static

graph, but the time-ordered sequence of these time-aggregated graphs models the dynamics of the whole network. An illustration of an exemplary temporal social network is presented in Figure 2.2 (this figure was originally introduced in Saganowski et al. [2014]).

Definition 4 does not say that the periods $T_1, ..., T_p, ..., T_K$ have to be of equal length, overlapping or non-overlapping or even whether they have to cover the whole ES. The only condition that has to be satisfied is that $t_1 \leqslant ... \leqslant t_p \leqslant ... \leqslant t_K$ and $t'_1 \leqslant ... \leqslant t'_p \leqslant ... \leqslant t'_K$, i.e. that the beginnings and the ends of the periods have to be in non-descending order. Moreover, as the TES defines two closure types for the beginning of the period and the end of the period, there are four types of periods allowed, see Definition 3.

The most typical variants of TSN are presented in Figure 2.3. Naturally, the choice of how to build a TSN depends on the goal one would like to achieve. As it was presented in Saganowski et al. [2012], the size and type of time windows may influence the results of the analysis of dynamic processes. In this study researchers analysed the group evolution in social networks, but it may also be the similar case for other types of processes. This is why the decision on how to create a TSN based on ES is of great importance. Still, a number of TSN variants seems to be quite natural when thinking about modelling temporal social networks. They are briefly described below.

Non-overlapping consecutive time-windows

The most obvious scenario is presented in Figure 2.3a. Here the time windows are non-overlapping and the whole period of ES is covered by neighbouring social networks following each other. The most important property of this approach is that each event from ES was considered just once, so there is no overlap in particular time windows.

Partially overlapping time-windows

Another way of creating TSNs results with partially overlapping time windows (Figure 2.3b). The reason to create such networks is that there is no strong "cut" that splits social networks. In fact, the border of time windows is a quite a sensitive place where many unique points may be lost, so this is the way to soften this border. One drawback is that here some events will

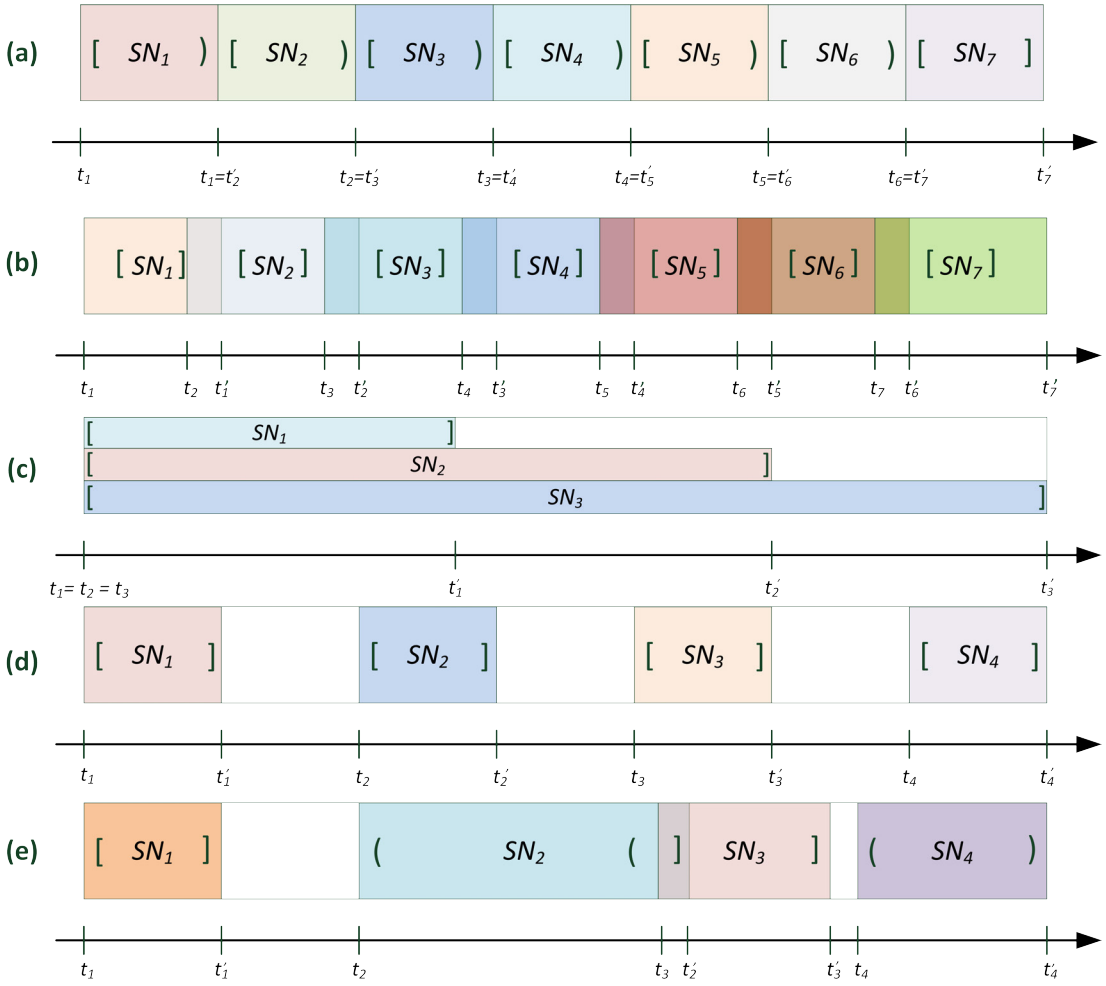


Figure 2.3: Different variants of Temporal Social Networks.

be taken twice, so the edge weights will be also affected.

Increasing time windows

This is a unique approach which, for every social network, uses the events from the beginning of the event sequence, and each social network is built over longer periods (Figure 2.3c). It is used particularly to verify how stable the network is, i.e. whether the structural measures change proportionally to the length of the period. On the other hand, it may be used when the network grows in terms of diameter. This way of building networks may also reveal when some structural holes appear. However, a drawback is most of

the burstiness (Barabási [2010]) becomes flattened.

Non-consecutive time windows

In the case of datasets with regular patterns it is often considered whether some periods shouldn't be skipped in order not to disturb this pattern (Figure 2.3d). For instance, if corporate e-mails are analysed and the goal of the analysis is to observe how the network changes on working days, weekends may be omitted. Naturally, a drawback is that some events are not included in the resulting TSN, but in some cases it may be an advantage rather than a loss.

Hybrid approaches

The last approach presented is the mixed one - some periods were longer, some shorter, some overlapping, some not (Figure 2.3e). Basically, the intention of showing this variant here is to show that the process of building the TSN does not necessarily have to be ordered or regular. Sometimes, when trying to capture some special events, it might be reasonable to build a mixed-mode network.

In this dissertation it was decided to build TSNs with time windows of equal size and non-overlapping, where each window is neighbouring with the next one - the case presented in Figure 2.3a. The argument is that each event from ES should be included just in a single SN, as we do not want to lose any event. More details on how the TSNs were created are presented in Chapter 5.

2.6 Discussion

Social networks are created upon human activities. Despite the fact that we are predictable to some extent, the social networks we extract are in fact dynamic. The factors behind the dynamics may be of differing natures, such as meeting new people, changing attitude towards others, changing jobs, moving from one place to another and so on. In fact, the most accurate method of representing human communication is the precise information about who contacted whom at which time - Event Sequence (see Section 2.2). By having this data, it is possible

to track how information or influence may potentially flow from one to another. However, without knowing about the content or the essence of the communication it is possible to make just some assumptions about potential paths of information diffusion. Moreover, as a single contact mostly contains just two entities: sender and recipient (directed case, as defined in this dissertation) or contacting parties (undirected case), this granularity cannot benefit from the mature apparatus of Social Network Analysis (SNA). It also requires a lot of storage in comparison to another approach - building a social network from ES, as defined in Section 2.3.1. In such case, the information about communication gets aggregated and the intensity of contacts is most typically expressed as weights over edges (events) between nodes (individuals), as presented in Barrat et al. [2004], Michalski et al. [2011a] or Michalski et al. [2012a]. This approach allows us to obtain a broader view of interconnections in networks and enables us to distinguish groups, hubs, nodes on boundaries of the network and lets us perform other analyses that are offered by SNA techniques (Carrington et al. [2005]; Kazienko et al. [2011]). Unfortunately, as the time-aggregated view of the network is used, all the events are of the same importance, without taking care of their timestamp. From an information or influence propagation point of view, the most important aspect is missing - the order of contacts. As it was stated in Pfitzner et al. [2013], in these networks one assumes transitive paths, which of course do not keep in event sequences or temporal social networks. Moreover, as the contacts within social networks are often *bursty* (Barabási [2010]), the static representation of networks will also ignore this fact leading to wrong conclusions about the dynamic processes taking place in it. This is especially crucial when modelling the spread of epidemics, since the accuracy of predictions may strongly influence the potential actions in healthcare (Masuda and Holme [2013]). To prevent the loss the temporal information, researchers more and more often use temporal representation of networks, and a comprehensive overview of methods of building temporal networks may be found in Holme and Saramäki [2012]. In this work, authors state that the literature on static social networks is many times larger than on temporal ones. This is for a natural reason: it is much easier to analyse time-aggregated networks, especially analytically.

Temporal Social Network, as defined in Section 2.5, is a trade-off between

the SN representation and contact sequence representation as defined in Holme and Saramäki [2012]. The author of this thesis agrees with authors of the above cited work that TSN representation may miss some unique points of interaction, but there are also some advantages. Firstly, as the research on temporal social networks is in its early stage, it is hard to name established temporal versions of structural measures. Secondly, the most granular representation also increases computational complexity, since each edge has to be treated individually instead of being aggregated to some extent within the social network - a part of TSN. Lastly, it is far easier to compare the social networks than the contact sequences when trying to answer the question: how do they change in time? Based on this argument, in this thesis temporal social networks are represented as introduced in Definition 4.

2.7 Summary

In this chapter four important concepts that will be used later in this thesis were introduced. Event Sequence ES is an event log which may be treated as a raw data indicating who contacted whom at which time. Then, by using ES a Social Network SN was defined as a time-aggregated static network connecting initiators and recipients of these events, built from nodes and edges. Since the goal was to take advantage of temporal aspects, by defining Time-limited Event Sequence TES (see Definition 3) it was possible to limit the ES to some given period. Finally, Temporal Social Network TSN is a time-ordered sequence of SNs that enables the use of typical SNA techniques in each SN. Temporal aspects are preserved by the appropriate use of TESes. With the introduction of temporal social network, the dynamic aspects of the underlying network layer may be represented. Now, it is time to focus on dynamic processes that take place in those networks.

In the next chapter, a number of those dynamic processes that occur everyday are introduced. Starting with the *diffusion of information*, a process of propagating information through the network, the *diffusion of innovations* is presented. It is a different concept, since it introduces a social change of an individual towards an idea or product. Lastly, *social influence* is described. Social influence is

a complex psychological phenomenon which concerns individuals facing a different opinion to their own. The question is whether they will keep their opinion or become influenced by others. Social influence is the process that will be a base for dynamic process taking place in the temporal social network. This dissertation will evaluate whether it is possible to maximize the spread of influence in the temporal environment.

Chapter 3

Diffusion Processes and Influence in Social Networks

3.1 Introduction

As we are immersed in the information era, we cannot avoid contact with information. We receive it from our relatives, newspapers, TV, colleagues and, naturally, from the Internet. The level of exposure to information is as never before, sometimes making us confused. But as the human being is a flexible phenomenon, we start to get used to this information stream. Nevertheless, it is worth analysing how information flows across the social networks and which particular processes are responsible for transmitting it and, sometimes, for changing our mind in some areas.

This chapter aims to present this domain from three different perspectives in order to show the multidimensionality of the problem. Starting from the idea of *information diffusion*, that is the process of information flow in social networks. Second important phenomenon is presented - *diffusion of innovations*. It is the theory of trying to explain why particular ideas have the chance to be adopted while the others do not succeed. This theory studies the process of personal adoption in detail, whereas the last part of this chapter introduces the theory of *social influence* and models of *social influence*. Spread of influence takes a different perspective from the former approach - by using models of social influence at a

local network level, the total effect of the process is observed. To conclude the chapter, three approaches are compared synthetically, in an attempt to allow the reader to better understand the similarities and differences between them.

3.2 Diffusion of Information

As it was stated in the abstract of this thesis, people do not live in isolation. It results in the conclusion that by being a part of multiple social networks, we are at the same time the recipients and transmission medium of information. Moreover, apart from our social circles, we may also receive information from external media, such as newspapers, Internet or television, which are sometimes called out-of-network sources (Myers et al. [2012]). This process is symbolically presented in Figure 3.1, where members of a small social network exchange some information (marked as red dotted arrows) over links previously established in the network. Apart from that external sources also spread this information. Each transmission of information is timestamped, since this is a time-respecting process.

Here, the term *information* represents an abstract that may be a simple rumour, it may extend somebody's knowledge, build an opinion on a particular subject or dramatically change someone's life. The process of diffusion of information focuses not on the content of this information, but on the technical capabilities of spreading this information across the social network or, more generally, the society. Trying to observe this process may be extremely challenging task, since it is impossible to have the full knowledge about social network structure, its dynamics and external sources. Moreover, depending on our own perception of the importance of that information, we may initiate contacts to inform our neighbours about it or to pass it when we meet people on different occasion. This distinction is heavily subjective and hard to model. That is why research in this area diverges in a number of directions summarized below. The state-of-the-art in this domain is presented and analysed in Barrat et al. [2008]. In this thesis the concepts of diffusion of information and diffusion of innovations are intentionally distinguished. The reason is that not all information may lead to changing on opinion on something. Nor does all information contain some idea. Some information may simply spread across the network without making any change in its

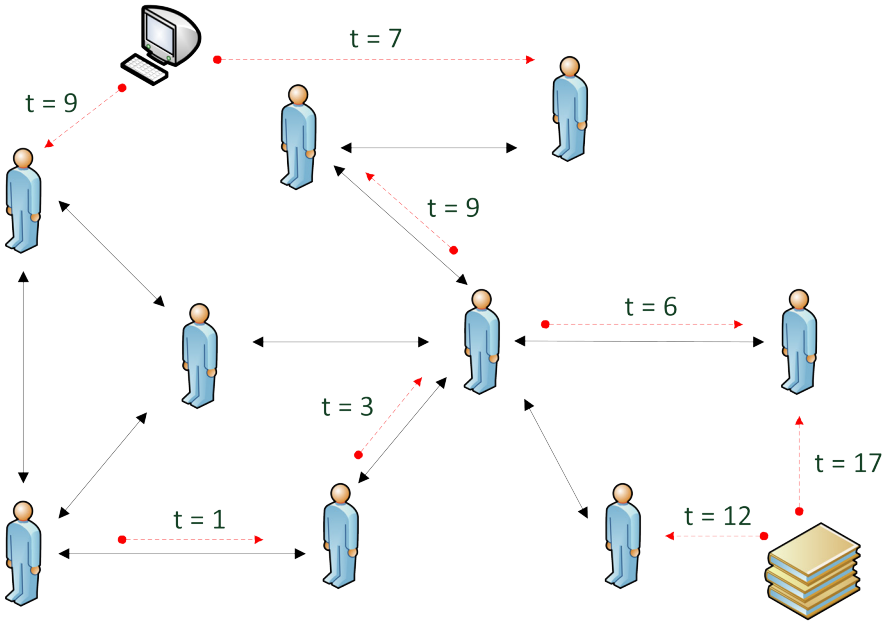


Figure 3.1: The information diffusion occurring in a social network with out-of-network sources (Internet and books). Black arrows represent the underlying network structure and red ones represent the diffusion of information in particular time moments.

state. In contrast, diffusion of innovations is a process where an idea, concept, attitude traverses through the network and changes its members in some way. Nonetheless, some researchers use both terms alternately.

3.2.1 Empirical Research

Researchers attempted to answer the question of whether the networks could be relied upon to transmit information and how long would it take. The most recognizable research is the work by Travers and Milgram [1969] which led to the definition of the small world model by Watts and Strogatz [1998]. This research unveiled that the real world social network has a relatively small average path length, which can be used to predict the speed of information diffusion and its ability to reach different segments of the network. Based on other data, similar evaluations were made of different content and environments, such as: chain-letters (Liben-Nowell and Kleinberg [2008]), blogosphere (Gruhl et al. [2004]), emails (Dodds

et al. [2003]) or twitter posts (Yang and Leskovec [2010]). It should be emphasized that a huge acceleration in research in this domain was possible thanks to the development of technology, which allowed the tracking of information flows more precisely. Still, as every social network differs, other results are obtained for each kind of network. Even if the nodes in the network remain the same, the variety of communication methods used to pass the information may also speed up or slow down the diffusion process. To consider this diversity, researchers developed another viewpoint on networks by defining a separate layer of the network for each communication method. This approach is called multi-layered (Jankowski et al. [2013a]; Kazienko et al. [2010]; Magnani and Rossi [2011]) or multiplex networks (Lee et al. [2012]). This kind of network helps in understanding how people are connected by using different mediums or how the edges and communities on different layers overlap. Studying the diffusion in multi-layered networks is more complex, since it requires more data sources to be captured and analysed. For instance, tracking e-mail and mobile communication requires obtaining the data from SMTP servers and telecoms. This is why research in the area is more oriented on simulations rather than on real world experiments (Gomez et al. [2013]; Michalski et al. [2013]).

3.2.2 Models of Diffusion of Information

Another direction in the area of diffusion of information is focused on modelling these processes. An extensive work summarizing the state-of-the art in this area (Bartholomew and Bartholomew [1967]) presents different models for different sociological phenomena and it is a good starting point for discovering the sociological background of particular models. Most of the presented models follow the Markov property (Markov [1951]), some suitable for discrete time and some others for continuous time. The most important from this dissertation's point of view are:

Birth-death processes

Birth-death processes are considered a special case of Markov process, where transactions from one state S_n are permitted only to neighbouring states S_{n-1}, S_n, S_{n+1} (Garcia [1990]). The etymology of the name is straightfor-

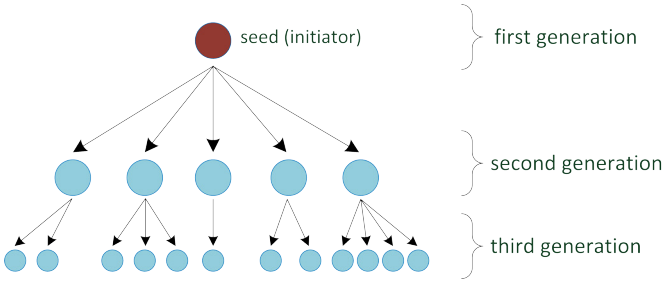


Figure 3.2: The diffusion of viral content represented as a branching process.

ward - these processes may be used to model the population, where people are born and die and the state S_n represents the current population size, λ_n the average birth rate and μ_n the average death rate. In the area of information diffusion these processes may be used for modelling the number of social network members knowing some information but also taking into account the possibility that someone may forget it. Moreover, to consider a process as a birth-death one, it is not realistic to assume that two people may obtain or forget the information at exactly the same time.

Branching processes

Branching processes have Markov properties, but the formulation of the problem is slightly different than in the birth-death problem (Kendall [1966]). Here, it is assumed that the process is parametrized by just one variable μ defining the number of expected children of an individual. It allows the calculation of the expected size of the nth generation, which is μ^n. These processes are widely used to model viral campaigns in social networks (Iribarren and Moro [2011]; Jankowski et al. [2012b]), especially while attempting to discover how "deeply" into the network structure information will spread if the number of neighbours an individual will infect is expressed as μ. An illustration of a viral content diffusion process represented as a branching process is presented in Figure 3.2.

Epidemic models

A novel direction in the modelling of information diffusion has its roots in modelling epidemics. Formal models that are most widely used for predicting disease spread in society are used more and more often to try to

model the diffusion of information. The most recognizable ways for modelling the epidemics (see Bailey et al. [1975] as a good reference), such as *Susceptible-Infectious-Susceptible* (SIS) or *Susceptible-Infectious-Recovered-Susceptible* (SIRS) are increasingly often adopted or even directly used in this area. Indeed, some further examination of these models on real world data shows that in some situations they may accurately describe the diffusion of information in social networks (Gruhl et al. [2004]; Xu and Liu [2010]).

There has also been an extensive research conducted in the area of rumour spreading, where researchers try to examine different models that sometimes extend the models presented above, but are more adjusted to model human behaviour (Dietz [1967]; Galam [2003]; Lefevre and Picard [1994]; Moreno et al. [2004]; Nekovee et al. [2007]; Trpevski et al. [2010]).

Each of the above models may be used as a way of modelling the information diffusion in the social network with any kind of information transmitted. There are also rare cases when researchers were able to evaluate models against real world data, see Apolloni et al. [2009], Yang and Leskovec [2010], and Zbieg et al. [2012]. It shall be noticed that these models do not assume that an individual will be somehow personally attracted by the information - his or her personal state is irrelevant from the perspective of the whole process. This is the major difference when comparing the diffusion of information against the innovation diffusion or social influence, where the attitude or persuasion of an individual implies the state of neighbouring nodes. This and other differences will be described in Section 3.5.

3.2.3 The Role of Individuals in Information Diffusion

However, it should be remembered that apart from out-from-network sources, the information is transmitted (or not) by the decisions of individuals. They become familiar with information, rate the quality, importance and relevance of it and decide whether to inform the others or not. So another branch of research in this area is devoted to studying the information diffusion process at the lowest level - the motivations of people to become information spreaders or keepers. A very extensive work of Palloni [1998] summarises and justifies the sociological advances

in this area, presenting also additional models of diffusion. Since the goal of this dissertation differs from presenting motivations of individuals in information diffusion, readers are advised to start with this survey if they are interested in extending their knowledge in this area.

3.3 Diffusion of Innovations

As it was already mentioned in Section 3.2, two concepts: diffusion of information and diffusion of innovations are distinguished. In the most cited book about diffusion of innovations, Everett M. Rogers states that *„Diffusion is the process by which an innovation is communicated through certain channels among the members of a social system. It is a special type of communication, in that the messages are concerned with new ideas"* (Rogers [2010]). This definition slightly contrasts with the point of view of the author, but for the purpose of this work it should be assumed, that the diffusion of information is represented as communication by Everett's definition and diffusion of innovations is the process of diffusion.

This definition leads to the conclusion that the diffusion of innovations among society results in social change of some type. To properly understand the idea of innovation, it should be remembered that the innovation itself must supersede some previous idea, i.e. it has some advantages of different kinds - economic, social, technological - that give it a chance to become accepted by an individual or a group of people. So, the innovation introduces a social change and the way this change occurs in an individual is presented in the next section. Here, in contrast to the previous section, which omitted the role of individuals in diffusion of information, it was decided to present the individual's perspective, since this aspect is more relevant to this dissertation.

3.3.1 The Innovation-Decision Process

Rogers [2010] presented a model of stages in the innovation-decision process. This process describes how a person becomes familiar with innovation, then how he or she makes the decision to adopt or reject it and, if adopted, how this person moves to the confirmation stage. It is an interesting study showing that this

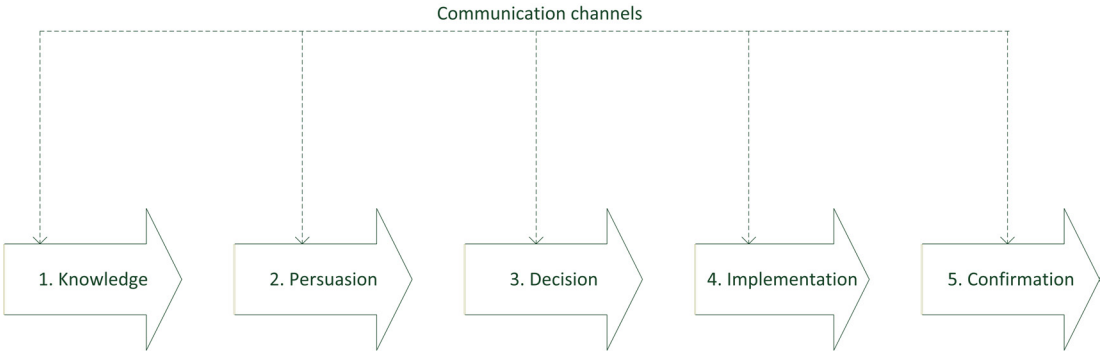


Figure 3.3: The model of innovation-decision process (Rogers [2010]).

process is complex and it involves many psychological and sociological aspects. The illustration of the model is presented in Figure 3.3 and each stage of it is described briefly below based on Rogers [2010].

1. Knowledge

The first stage, the knowledge stage, happens when an individual is exposed to the innovation's existence and gains some understanding of how it functions. It is often assumed, that this process is passive, that is a person does not actively look for a new innovation, but becomes informed about it by one of the communication channels (TV, peers, Internet). However, some researchers argue that this process requires activity, since the awareness of the innovation must be somehow initiated. An interesting part of this stage is also the question of how a person becomes aware of the need for this innovation. As the history of innovations shows, some innovations are the fulfilment of our needs, but some generate these needs. So the innovation may simply be better than its antecedent or, sometimes by being exposed to the innovation, we learn or discover new needs that have to be addressed and, indeed, these needs are addressed immediately by this innovation. The latter case is often described with the example of the Apple company. Their products often show to the customers some properties which instantly become highly desired. Here, the need for particular innovation is being born to become instantly fulfilled it by a particular Apple product (Carlson and Wilmot [2006]).

2. Persuasion

After obtaining knowledge of the innovation, a person aligns it to his or her viewpoint, forming either a positive or a negative attitude towards it. Compared to the previous stage, persuasion is more related to the feeling rather than to the knowing, since attitude is a subjective concept and by having the same knowledge two people may form a different opinion of the innovation. Of course, since the knowledge of an innovation is rarely total, the uncertainty about it may lead some people towards forming a positive attitude anyway. They believe that it may fulfil their needs, and some others will try to avoid the risk and will present a more conservative approach - rejecting the innovation. Still, the main outcome of this stage is either a favourable or unfavourable attitude toward the innovation.

3. Decision

As a person possesses the knowledge and the attitude towards the innovation, they either decide to adopt the innovation or to reject it in the decision stage. This part of the process is crucial from the social change point of view, since it is the moment when a person becomes convinced or not as to the innovation and performs some action. Rogers [2010] assumes that when a person decides to adopt the innovation, they also decide to make full use of it. On the other hand, rejection may be one of two kinds, *passive rejection* or *nonadoption*. The former means that the use of the innovation was never considered, and the *active rejection* (nonadoption) is the conscious decision not to adopt the innovation.

4. Implementation

After making a positive decision about to adopt the innovation, an individual puts the innovation into use and this is the first moment when the social change becomes visible to others who may then be exposed to it. Until this part of the model, an individual was facing a mental exercise rather than real action about how he or she actively adopts the innovation.

5. Confirmation

As Rogers [2010] states, some researchers finished the model on the *Im-*

plementation stage. However, it seems that innovation adopters often seek confirmation or reinforcement of their decision. They may even reverse their decision, if they become exposed to conflicting messages about the innovation. This is the stage where different kinds of dissonance may occur which should be eliminated or at least reduced. This dissonance comes from facing the attitudes of the closest neighbourhood towards the innovation adopted by an individual, from uncertainty about how to use the innovation or from other sources.

Last but not least, this model also involves *communication channels* by which different information is distributed. This information may be about the innovation itself, e.g. providing knowledge about it or about the adoption of the innovation by an individual or their peers. The communication channels are the same as the ones presented in Section 3.2, i.e. peers or out-of-network sources, such as mass media. These communication channels may be also considered as the sources of influence which affect individuals' decisions at every stage of the innovation-decision process.

3.3.2 Adopter Categories

The process of becoming an innovation adopter is time ordered, as it was shown in Section 3.3.1. Still, some stages may be shorter, the others longer, depending on the innovation type and the behaviour of an individual. Naturally, not all individuals adopt an innovation at the same time and this happens for various reasons. This differentiation led to the categorization of adopters which was also presented in Rogers [2010]. He enumerated five categories of adopters, referred to as *ideal types* and placed them on a Gaussian distribution plot, see Figure 3.4.

The understanding of these categories may be helpful in knowing the potential pathways of innovation across society, but it should be emphasised that this is an ideal categorization which may be not true for every innovation studied.

Innovators

The typical trait of innovators is venturesomeness. They are eager to learn about innovations, to try them, even if adopting the innovations will cost

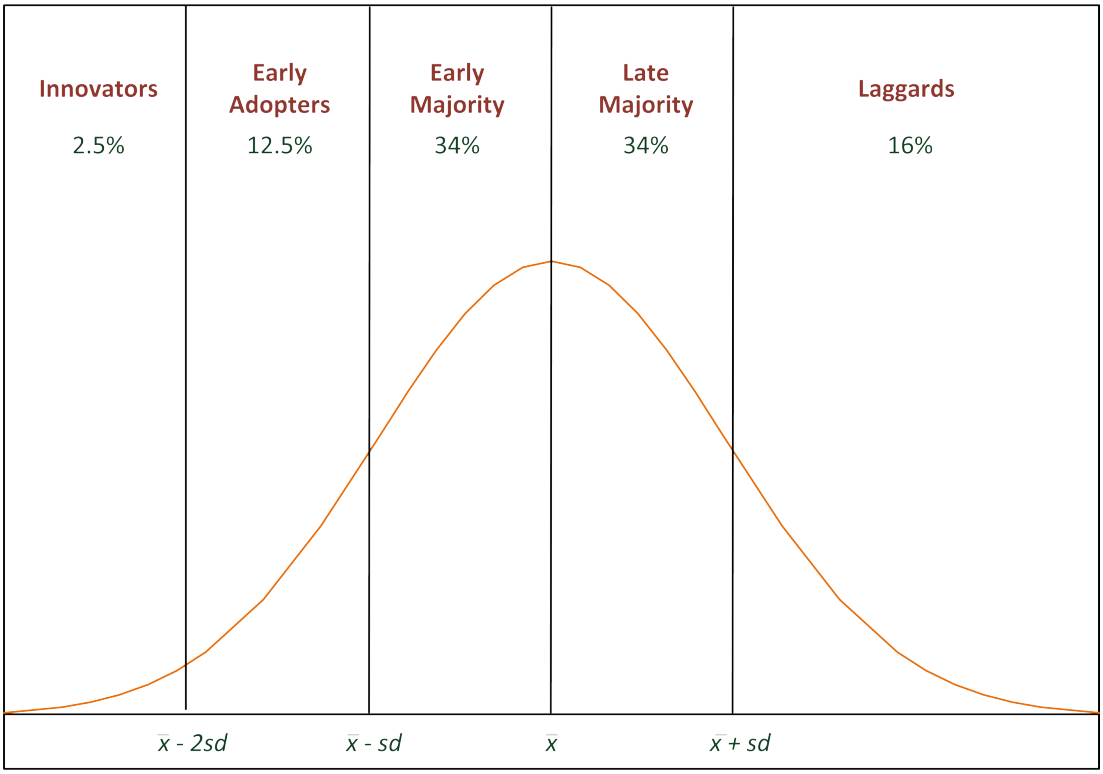


Figure 3.4: The *ideal types* representing adopter categories in the diffusion of innovations process (Rogers [2010]).

them more than an alternative, ans introduce risk. Despite the geographical distance, innovators form cliques that are communicating frequently. As Rogers [2010] states, they *play a gate-keeping role in the flow of new ideas into a social system.*

Early Adopters

Society needs people that are local to them to learn about innovations and this is the role of early adopters. They adapt the innovation and play the role of advisers in their environments with whom the others would like to discuss the potential of the innovation. Compared to innovators who may be described as cosmopolitans, early adopters are local leaders. They reduce the degree of uncertainty about innovations among their peers, playing the central role in their groups.

Early Majority

They adapt ideas just before the average member of the social system, but they do not hold a central position like early adopters, avoiding leadership.

Late Majority

The late majority are cautious about their decision and presents more often a sceptical opinion on the innovation. Members of this group more often require the motivation from others to be convinced as to the innovation, so the degree of uncertainty about the idea has to be reduced before they adopt it.

Laggards

Laggards represent cautiousness in adopting the innovation and this may be due to several factors. Their economic position may force them to avoid risky behaviour, before they can be sure that the innovation will become widespread. But some other reasons are also possible. For example, family traditions, or their position in the social network resulting in communication delay. It is often observed that when the laggard adopts one innovation, its successor is already adopted by innovators. There has been discussion as to whether the term *laggard* is not too pejorative when applied to this group, but no other has been agreed since the 1960s.

3.3.3 Models of Diffusion of Innovations

This subsection aims to present the three most popular mathematical models used to study the diffusion of innovations. They are presented in chronological order according to their introduction. An extended presentation of these and some other models can be found in Carrington et al. [2005], which also provides an extraordinary set of references to empirical studies on which these models were evaluated. These models may be divided into three categories:

- Macro models - assuming the perfect social mixing and no proximities between people.
- Spatial Autocorrelation - which reflect the physical distance.
- Network Models - focusing on the neighbourhood of an individual.

The first two models: the logistic and Bass model belong to macro models, whereas *Moran's I* belongs to spatial autocorrelation, and the last group covers the network models.

Parametric logistic model

This model, belonging to the group of macro models, is relatively simple, since it uses a one-parameter logistic function (Carrington et al. [2005]):

$$y_t = b_0 + \frac{1}{1 + e^{-b_1 t}}, \tag{3.1}$$

where y_t is the proportion of adopters to the whole population in time t, b_0 denotes the y intercept and b_1 is the parameter to be estimated. The simplicity of this model limited its applications, so it was substituted by the Bass model.

Bass Model

The Bass model is a two-parameter model that overcomes the obvious limitations of the one presented in Equation 3.1. It can be used to forecast expected reach of innovation or to estimate the diffusion rate. Its mathematical formulation is as follows (Bass [1969]):

$$y_t = b_0 + (b_1 - b_0)Y_{t-1} - b_1(Y_{t-1})^2, \tag{3.2}$$

where y_t is the proportion of adopters to the population in time t, b_0 is a rate parameter for innovation, Y_{t-1} is the number of adopters prior to the time t and a new parameter b_1 represents the rate parameter for imitation, i.e. the degree of adoption due to prior adopters. The most important improvement is that this model allows the incorporation of the percentage of adopters at each time point. The Bass model is widely used in empirical studies (e.g. Bass et al. [1994]; Dodds [1973]; Tigert and Farivar [1981]; Wright et al. [1997]) as well as often being extended to cover particular phenomena (Mahajan and Muller [1996]; Tseng and Hu [2009]).

Moran's I

Moran's I is an early model which tests spatial association, geographic clustering of adoption and it is based on the fact that it is relatively simple to obtain proximity data. The model is formulated as follows (Moran [1950]):

$$I = \frac{N \sum_i^N \sum_j^N d_{ij}(y_i - \overline{y})(y_j - \overline{y})}{S \sum_i^N (y_i - \overline{y})^2}, \tag{3.3}$$

where I indicates the level of autocorrelation, i.e. the level of clustering, N is the sample size, D represents the distance matrix and $d_{ij} \in D$, adoption is indicated by y, its mean value is $\overline{y}$, and S is the sum of all distances in the distance matrix. This approach measures the degree to which nodes that are connected to one another deviate from the average behaviour in the network similarly or differently (see Moran [1950] and the description in Carrington et al. [2005]). The main drawback of this approach is the lack of usage of network properties, i.e. it ignores the underlying network structure behind the diffusion process. That makes the Moran's I in some sense similar to the macro models, since none of them consider the actual network. The last approach tries to overcome this limitation by being strictly dependent on the network.

Network Models

The network-based approach uses a different point of view. Instead of looking from the global perspective at the diffusion process among society or the group, it analyses the ego (actor) networks (Prell [2011]; Valente [1996a,b])

to calculate the probability of adopting the innovation. It calculates the network exposure of an individual as the fraction of neighbours that adopted the innovation (Carrington et al. [2005]):

$$E_i = \frac{\sum w_{ij} y}{\sum w_i}, \tag{3.4}$$

where w is the weight of the network edge from node i to j and y is the vector of adoptions. This approach is the baseline for all the network-dependent models which incorporate the network structure for studying diffusion. Still, by being a very low-level one, it requires to know the topology of the network and obtaining this topology may be a challenging task.

Three different approaches for modelling the diffusion of innovations were presented in this section. They differ by point of view: the macro level attempts to estimate the outcome of the process without focusing on local properties of individuals, *Moran's I* uses the proximities and the network models represent the bottom-up approach starting from the node's perspective to end up with the global proportion of adopters.

These models may be applied to different situations, since it is unlikely that the researcher will have full knowledge of the environment where the diffusion process will take place. So when attempting to predict the outcome of the diffusion of innovations process, which model may fit the best must be known, depending on the knowledge of the innovation type and the social structure that adopts this innovation.

3.3.4 Summary

The above introduction to the process of diffusion of innovations plays an important role in this chapter. It shows how character, location in the social network and other aspects of an individual place him or her in the diffusion process. Moreover, the complexity of the individual decisions leading either to adoption or rejection an innovation is also presented. Since this was just an introduction to the far richer area combining psychology and sociology, this chapter now turns in the direction of the essence of the whole dissertation - the social influence which

is a theory studying why people become influenced. It is definitely broader than the diffusion of information and presents a different approach than diffusion of innovations.

Concluding the topic of diffusion of innovations and just to demonstrate to the reader that it may be even more complex, some other competing approaches are mentioned. They were actively studied in Jensen [1983], Dunn and Gallego [2010] and Venkatraman et al. [1994].

3.4 Social Influence

3.4.1 Introduction

Social influence is defined as change in a person's cognition, attitude, or behaviour, which has its origin in another person or group (Raven [1964]). By using this definition it is relatively easy to find the basic difference with the diffusion of innovations. Here the major role in an individual's change is played by their neighbourhood whereas in the diffusion of innovations the main cause of social change was the innovation or product itself. Still, in most cases it is hard to distinguish what were the most important factors when adopting some innovation or attitude - peers, out-of-network sources or personal reasons. But when trying to compare these two concepts, the essential assumption is that in social influence the neighbourhood plays a more important role in adopting some behaviour than the subject, while diffusion of innovations puts the subject at the front of the decision process.

In this section the concept of social influence from the sociological perspective is presented. It is compared against diffusion of information and innovations to give the reader a full understanding of the rationale behind the differentiation of these concepts. It also introduces the models of influence; one of them will then be used in the research part of this thesis.

3.4.2 Sociological Background

Social influence is one of the major research areas in social psychology. It is a scientific study devoted to the observation of how people change each other's behaviour and attitudes. Another good definition of social influence apart from the one quoted in Section 3.4.1 appeared in Aronson [2003]: *„Social influence is the effect that people have upon the beliefs or behaviours of others"*. It is obvious that being a social species we are exposed to external influence, which ideally spreads among groups. On the other hand, some other phenomena, especially *homophily*, may be also important explanation why groups are somewhat similar (McPherson et al. [2001]). The influence may be of different kind: intentional (e.g. rewards or punishments) or unintentional (being a person considered important in a society). However the question is how people react to the social influence of others (Turner [1991]). Aronson distinguished three kinds of responses to social influence which better define why we tend to conform when exposed to the influence. These are: *compliance*, *identification* and *internalization*; they are briefly described below (Bagozzi and Lee [2002]; Kelman [1958, 1961]). These three types of reasoning behind adopting someone's point of view try to cover all the possible motivations of a person becoming influenced.

Compliance

This is the behaviour that makes human beings no different to other animals, since by behaving compliantly they are trying to gain some reward or avoid punishment. It is most often observed that this behaviour lasts as long as the promise of the reward or the threat. It is very unlikely to become a habit if these external motivations disappear (Cialdini and Goldstein [2004]) and a person has just the slightest or even no conviction at all that the idea of the influencer is right from this person's point of view.

Identification

Identification also represents a motivation in which it is unlikely that the influenced person is fully convinced of the idea of influencer. Instead, he or she is trying to become like the influencer, believing that this is the desired behaviour. Still, this phenomenon is likely to have a long-lasting

effect, which differentiates it from *compliance*. So, the person starting to behave or think similarly to the influencer is most likely not intrinsically satisfied. The true reason is that it may lead to the positive relationship with the person or group (Aronson [2003]; Mugny et al. [1984]).

Internalization

Last but not least *internalization* is the most permanent response to social influence. Here the influenced person believes that the opinion, behaviour or attitude of an influencer is intrinsically right, so he or she accepts it internally, extending or modifying their own system of values. This is a true identification with the influencer, arising from the individual's commitment to the subject rather than from trying to adapt to someone else behaviour, gain rewards or avoid punishment. Moreover, due to this strong commitment, this person may even advocate this value to others, which is unlikely in the previously presented reactions (Lepper [1983]; McCauley [1989]; Ryan and Stiller [1991]).

These three responses to social influence are most likely met in human behaviour. However, sometimes it may happen that there is no single response type found, but adopting some value from others may be due to a combination of the above. Nevertheless, most likely, there is one crucial reason why someone is successfully influenced, the others being only side effects of it.

From the influencer perspective, it may be important to strengthen the effect of influence and different researchers try to provide methods of how this may be done. For instance, three factors increasing the likelihood of becoming influenced that are the part of *social impact theory* are provided in Latane [1981]:

- *Strength* - the importance of the influencer to the individual.
- *Immediacy* - physical and temporal distance of the influencer to the individual.
- *Number* - the size of the group influencing an individual.

Another work attempting to study the phenomenon of increasing the influence effect (Cialdini [2001]) recognizes the following techniques:

- *Reciprocity*, i.e. willingness to return a favour.
- *Commitment and Consistency* - being committed to the system of values.
- *Social Proof* - the need of being committed to the change by observing others behaviour.
- *Authority* - obeying authorities.
- *Liking* - willingness to follow the argumentation of people an individual likes.
- *Scarcity* - limiting resources may force people to act differently than they would without limited resources.

Of course, the above introduction to sociological background of social influence may be considered as limited, but it is not the main goal of this dissertation to provide all the reasons why people become influenced and how to maximize the influence in terms of sociology.

Now this chapter turns to presenting the most commonly used mathematical concepts for modelling social influence. Their goal is to model the process and at the end, to estimate the overall number influenced by making some prior assumptions taken from the sociological background.

3.4.3 Models of Social Influence

3.4.3.1 Introduction

Before presenting the most popular models of social influence in social networks, one sentence from Hedström and Bearman [2009] by D.J. Watts and P. Dodds may be mentioned: *„(...) it is still the case that formal models of social influence suffer from a dearth of realistic psychological assumptions"*. The problems of fitting real world data to models and trying to answer the question whether a particular influence process may be modelled with a selected approach still challenges researchers. The problem lies in the complexity of human behaviour and the impossibility of separating social processes that are occurring simultaneously. Still, many results achieved in this area tend to contradict this pessimistic point

of view of Watts and Dodds. The continuous development of models or models' variations suggests that models will fit the reality much better in next few years (e.g. Aral and Walker [2012]). On the other hand, there still remains the gap between models and psychology that requires to be intensively studied to find the psychological rationale of particular behaviour expressed in the model.

The presented models of social influence may look similar to the network models presented in Section 3.3.3, since they focus on a network structure to answer the question whether someone will become influenced or not. This similarity is not accidental, since the point of view is identical. However, due to the fact that becoming influenced does not always mean adopting an innovation, different psychological processes may play an important role here. Moreover, social influence requires an external individual (an influencer). Because of those two reasons, in this thesis these two terms are distinguished. Further arguments are presented in Section 3.5.

Since the strength of social influence depends on many factors such as the strength of relationships between people in the networks, the network distance between users, temporal effects, characteristics of networks and individuals in the network (Sun and Tang [2011]), it is relatively hard to model all these factors combined. However, some research shows that under some assumptions there exist models that fit the reality well (Marsden and Friedkin [1993]; Masuda and Holme [2013]; Robins et al. [2001]). The main models that are most commonly used in this area are as follows:

- *The Linear Threshold model* (LT),
- *The Independent Cascade model* (IC),
- *The Voter Model* (VM),
- *The Naming Game* (NG).

Each of them tries to consider the psychological background, but as it was previously stated, sometimes it is just a loose interpretation of humans' behaviour, that still fits the reality. For these models their recent variants which are suitable for some real world cases are also presented.

From the historical perspective, studying the social influence in terms of analytical process was the case of trying to model how influence spreads in time. Starting from a set of influenced nodes in time t_0 which are in this work denoted as $\Phi(0)$, as time unfolds, more and more neighbours of $\Phi(0)$ become influenced if they fulfil the model criteria. Most typically, these processes are modelled in directed graphs and focus on a *progressive* case, where nodes may become *influenced* from an *uninfluenced* state, not the other way round (Kempe et al. [2003]). Since this is a network approach, the influence process runs through edges in the graph and most typically no other external factors of influence are considered, such as out-of-network sources.

3.4.3.2 The Linear Threshold Model

The most recognizable model for social influence is Granovetter's Linear Threshold Model (Granovetter [1978]), but a similar approach was also proposed by Schelling [1978]. In this model, a node v is influenced by each v's neighbour w according to a weight $b_{v,w}$, such that $\sum_{w \in N_v^{inf}} b_{v,w} \leqslant 1$. Then each node v chooses a *threshold* θ_v from the interval $[0, 1]$ and this threshold represents the value which has to be overcome by the aggregation of v neighbours' influence in order to influence the node v. So the formal condition of influencing the node v is as follows:

$$inf(v) = \begin{cases} 1, & \text{when} \quad \sum_{w \in N_v^{inf}} b_{v,w} \geqslant \theta_v \\ 0, & \text{when} \quad \sum_{w \in N_v^{inf}} b_{v,w} < \theta_v, \end{cases} \tag{3.5}$$

where $inf(v)$ is the result of influence process for node v, θ_v denotes the threshold level for node v, N_v^{inf} the set of influenced neighbours of node v and $b_{v,w}$ represents the influence weight of node w on node v. The influence process ends when no more nodes can be influenced - this is the stop condition.

In this case, the value of θ_v represents the individual's chances of becoming influenced when enough of its neighbours are influenced. So all the psychological factors are included in this parameter and it should be also underlined that this approach represents the individual's perspective rather than the influencer perspective. Granovetter illustrated the model with the hypothetical case of a riot.

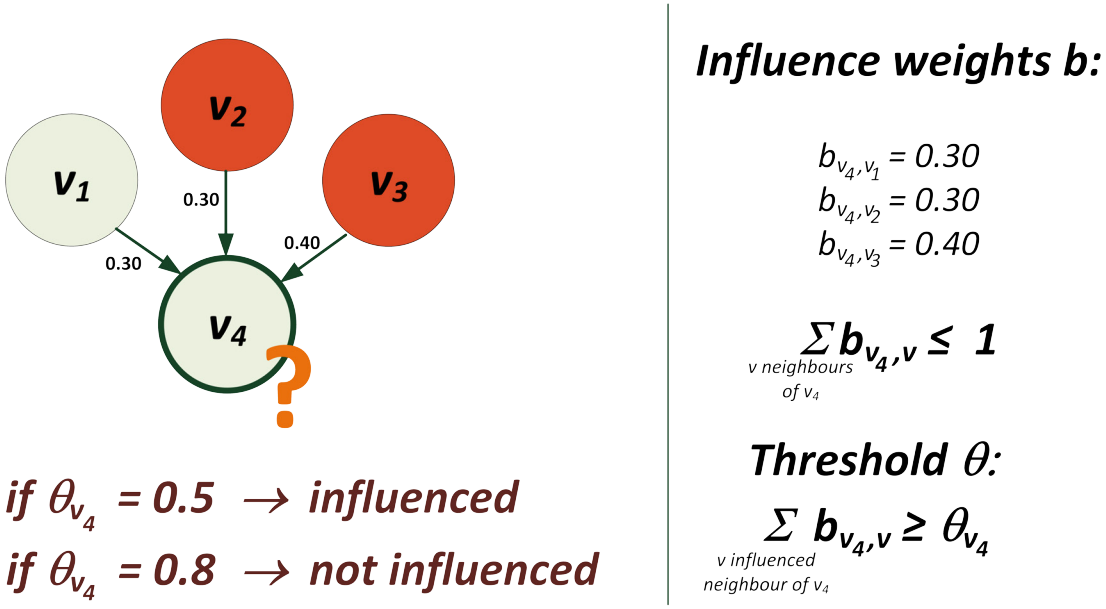


Figure 3.5: An illustration of the Linear Threshold model.

Since individuals were unsure what the costs and benefits of joining it are, they observed their peers and considered joining only when enough of their neighbours joined the riot. Otherwise, they refrained.

Of course, the greatest question is how to assign particular values of θ to individual nodes. There are two most typical approaches: draw them from a probability distribution $f(\theta)$ (Granovetter [1978]) or hard-wiring them at a fixed value (Berger [2001]; Peleg [1997]). The most interesting and realistic scenario is the former, i.e. drawing θ_v from a distribution, since the distribution represents both the average tendencies and also the heterogeneity present in the population. Lowering or raising the mean of $f(\theta)$ would modify the general susceptibility of the population, while increasing or decreasing the variance would correspond to an increase or decrease in variability in susceptibility among individuals (Hedström and Bearman [2009]). Still, hard-wired thresholds are also often considered in science. An illustration showing how the LT model works is presented in Figure 3.5 and the model is formalized as Algorithm 1 (based on Zafarani et al. [2014]).

The LT model became a core of many modifications or extensions. For instance, Goyal et al. [2010] extended this model by introducing temporal decay, as well as factors such as the influence-ability of a specific user, and influence-

Algorithm 1 Linear Threshold model

Require: Graph $G(V, E)$, set of initially influenced nodes $\Phi(t_0)$, thresholds θ_v, influence weights $b_{v,w}$

return Final set of influenced nodes $\Phi(K)$
k = 0;
Uniformly assign random thresholds θ_v from the interval [0, 1];
while $k = 0$ or $\Phi(t_{k-1}) \neq \Phi(t_k)$ **do**
  $\Phi(t_{k+1}) = \Phi(t_k)$;
  $uninfluenced = V \setminus \Phi(t_k)$;
  for all $v \in uninfluenced$ **do**
    if $\sum_{w \text{ influenced neighbour of } v} b_{v,w} \geqslant \theta_v$ **then**
      influence v;
      $\Phi(t_{k+1}) = \Phi(t_{k+1}) \cup \{v\}$;
    end if
  end for
  $k = k + 1$;
end while
$\Phi(K) = \Phi(k)$;
Return $\Phi(K)$;

proneness of a certain action. On the other hand, Barbieri et al. [2013] proposed topic-aware extensions of the LT model. In Pathak et al. [2010] authors considered multiple cascades of the LT model and they allow nodes to switch between them, whereas Borodin et al. [2010] introduced a number of modifications to the competing model variant: they force nodes to draw one cascade they join at the end of the process or consider the mutual influence of cascades on each other.

3.4.3.3 The Independent Cascade Model

The next model has its roots in interacting particle systems (Durrett et al. [1988]; Liggett [1985]) and it is called the Independent Cascade model - IC (Goldenberg et al. [2001]; Kempe et al. [2003]; Król [2014]). Again, the process starts with a set of influenced nodes $\Phi(0)$, but each node v in the network has assigned a probability $p_{v,w}$. According to this probability the node v has a single chance to influence its neighbour w and if it fails, it will have no other chance. If it succeeds, w will become influenced in the next step. The process runs until no

more influences are possible. The IC model is presented as Algorithm 2 (based on Zafarani et al. [2014]).

Algorithm 2 Independent Cascade model

Require: Graph $G(V, E)$, set of initially influenced nodes $\Phi(t_0)$, activation probabilities $p_{v,w}$

return Final set of influenced nodes $\Phi(K)$
k = 0;
while $\Phi(t_k) \neq \{\}$ **do**
  $k = k + 1$;
  $\Phi(t_k) = \{\}$
  for all $v \in \Phi(t_{k-1})$ **do**
    for all w neighbour of v, $w \notin \cup_{j=0}^{k} \Phi(t_j)$ **do**
      $rand$ = generate a random number in [0,1];
      if $rand < p_{v,w}$ **then**
        influence w;
        $\Phi(t_k) = \Phi(t_k) \cup \{w\}$;
      end if
    end for
  end for
end while
$\Phi(K) = \cup_{j=0}^{k} \Phi(t_j)$;
Return $\Phi(K)$;

In the case of this model, from a psychological perspective, the influencer becomes more important, since he or she holds the probability $p_{v,w}$ and this is one of the differences between IC and LT models. In the LT model, the influence process parameter was assigned to the uninfluenced node and in the IC model it is held by the influencer. Just as in the previous model, the probability may be fixed or drawn from a distribution $f(p)$.

Again, there are many variants of the IC model. The already mentioned work by Barbieri et al. [2013] introduced the topic-aware approach also for this model, while Kempe et al. [2005] studied the *decreasing cascade model*. One of the problems with the base LT and IC models is that they need to assume the influence probabilities and there are works that try to obtain these probabilities from past propagations. It may not be considered as an extension of the base model, but an approach to make the probabilities or threshold more realistic. One of the works

in this area is Saito et al. [2008]. There also exists an approach to model multiple independent cascades in the network (Bharathi et al. [2007]).

3.4.3.4 The Voter Model and the Naming Game

An interesting case of influence in networks is the case where two separate opinions or influences are competing in society. This phenomenon may be observed in many situations and it has its roots in studying the consensus processes (Lu et al. [2009]) or the language dynamics (DallAsta et al. [2006]). Below there are two variants of the process presented: the Voter Model (VM) and the Naming Game (NG) .

The Voter Model introduced in Clifford and Sudbury [1973] and extensively analysed later in Holley and Liggett [1975] assumed that each node in the network can hold one of two opinions and by interacting with others it may switch the opinion to the opinion of the peer. The model also introduced the degree of conformity which defines whether a node will follow the majority (conformist) or minority (non-conformist), see Javarone [2014].

On the other hand, the Naming Game, also referred to as the binary-agreement model (Xie et al. [2011]) introduced another variant of forming the opinion or spreading the influence. At any time a node may possess one of two competing opinions or two opinions simultaneously. In a given step in time, we choose a node randomly, designate it as the speaker, choose one of its neighbours randomly and designate it as the listener. The speaker proceeds to convey its opinion to the listener (selected randomly if it possesses two). If the listener possesses this opinion already, both speaker and listener retain it while eliminating all other opinions; otherwise, the listener adds the opinion to his list (Xie et al. [2012]).

Both of these models are useful in studying common phenomena occurring in social networks which involves binary options, such as reaching the consensus on contradictory opinions or observing which of the competing parties will win the election. The current research trends suggest that these models will be actively studied and extended in the future (e.g. Jankowski et al. [2012a]; Li et al. [2013]; Maity et al. [2013]; Mobilia [2013]; Rogers and Gross [2013]; Zhang et al. [2013]).

3.4.3.5 Models of Influence - Summary

The above presented models are just a selection of models which allow the study of the influence process in networks analytically. As it was presented, they differ by the perspective (LT versus IC), and by the number of competing influences (VS and NG versus others), but all of them are linked to the same process - social influence. Sometimes their applicability is limited, but the empirical research shown, so far demonstrated that they model human behaviour accurately in some cases (Marsden and Friedkin [1993]; Robins et al. [2001]), even if the psychological background of an individual is more complex than just a single parameter.

At this point it is worth comparing the above presented phenomena dealing with the diffusion or social influence in social networks. It may help the reader to see the similarities and differences between them. This comparison is presented in Section 3.5. It will be followed by the global perspective - the spread of influence in social networks.

3.5 Comparison of Diffusion and Influence Processes

Table 3.1 lists all of the discussed social phenomena presented so far: diffusion of information, diffusion of innovations and social influence. They are briefly summarized to provide a quick overview of similarities and differences between them.

While summarizing these three approaches or theories, there are two questions that should be addressed. Firstly, why are there so many similarities between diffusion and innovations and social influence? In this dissertation diffusion of innovations is rather a process of becoming convinced and committed by the individual itself. This person is observing how his family or relatives react to a given innovation, whether they adopt it or not, what opinion about it is presented in mass media? This is why individual factors play the major role here. On the other hand, the process of social influence is mostly invoked by others and the individual eventually becomes influenced, or not. So diffusion of innovations is rather an internal process becoming convinced of some idea or innovation, while social influence introduces some pressure towards a person that is about to become

influenced.

Secondly, why do researchers use the term *diffusion* for two processes: diffusion of innovations and diffusion of innovations? If using the physical definition (Philibert [2005]), diffusion is *the movement of a substance particles* and there is no possibility that the number of particles will increase. So, when looking at this definition it is rather hard to justify the usage of this term, since while sharing some information we do not loose it. On the other hand, the process of diffusion assumes that there is *movement from the region of high concentration to a region of low concentration* and it might be key to proper understanding of the reasons for this name. Here, highly concentrated information or ideas (just a number of seeds or sources) diffuse towards regions with low concentration, reducing local maxima of density.

Table 3.1: Comparison of diffusion and influence processes in social networks.

	Diffusion of information	**Diffusion of innovations**	**Social influence**
Short description	General process of transmitting the information between individuals or between out-of-network sources and individuals. It includes any kind of information. The role of out-of-network sources is important.	This process reflects the diffusion of innovation or idea through the network.	Social influence is defined as change in a person's cognition, attitude, or behaviour, which has its origin in another person or group.
The content of transmission	Any information, regardless of the content.	Idea, attitude, innovation.	Since the process of social influence does not focus on the product or innovation, the content may be anything that can a man be influenced with.
Models	Birth-death processes, branching processes, epidemic models.	Parametric logistic model, Moran's I, Bass model, network models.	Linear threshold, independent cascades, naming game, voter model.
Application	Speeding up or controlling the information flow over the network, especially social media application.	Marketing products or promoting desired behaviour.	Politics, marketing, social behaviour, daily habits.

3.6 Spread of Influence in the Social Network

In Section 3.4, a number of properties of the social influence process were presented. Overall, social influence is the process where an individual becomes influenced by others to some idea, behaviour or attitude and the role of external entities is underlined here. However, as this process starts with a single influencer or a group of them, we can see that despite the fact that the influence itself is rather an individual case, it spreads among others at a network level. It means that it may be observed as a complex psychological and sociological process when thinking about the reasons why an individual is becoming influenced, but when taking another perspective the whole outcome of the influence process may be studied. This outcome is referred to as the *spread of influence*, i.e. the reach of the influence in a given network. It is most often measured by the number of influenced nodes after the process ends according to the fixed number of initially influenced individuals. The spread of influence for social network and temporal social network is defined formally below.

Definition 5. *The Spread of Influence* $SI^{SN}(SN, m, P^m, SC, \Phi(0))$ *for a social network* SN*, a given propagation model* m *and its parameters* P^m*, stop condition* SC *and initial seed set* $\Phi(0)$ *is the total number of influenced nodes from* SN *after the stop condition* SC *is reached:* $SI^{SN}(SN, m, P^m, SC, \Phi(0)) = |\Phi|$*. Here,* Φ *denotes the set of influenced nodes. Since the spread of influence* $SI^{SN}(SN, m, P^m, SC, \Phi(0))$ *is considered for a fixed propagation model* m*, its parameters* P^m *and stop condition* SC*, it will be further denoted as* $SI^{SN}(SN, \Phi(0))$*. To shorten the further textual content,* $SI^{SN}(SN, \Phi(0))$ *will be referred to as SI.*

Spread of influence for social networks is an iterative process, where for each iteration i it is possible that new nodes become influenced. The stop condition SC in Definition 5 refers to the definition of a given propagation model m. As it is presented in Section 3.4.3.2, the LT model reaches its stop condition after the iteration for which no more nodes can be influenced. The other propagation models may use different stop condition definitions, e.g. a fixed number of iterations.

Definition 6. *The Spread of Influence* $SI^{TSN^K}(TSN^K, m, P^m, p, \Phi(0))$ *for a temporal social network* TSN^K*, a given propagation model* m *and its parameters*

P^m, period T_p and initial seed set $\Phi(0)$ is the total number of influenced nodes $SI^{TSN^K}(TSN^K, m, P^m, p, \Phi(0)) = |\Phi(T_p)|$, where $\Phi(T_p)$ denotes the set of influenced nodes at the end of period T_p. Since the spread of influence $SI^{TSN^K}(TSN^K, m, P^m, p, \Phi(0))$ is considered for a fixed propagation model m, e.g. LT, it will be further denoted as $SI^{TSN^K}(TSN^K, p, \Phi(0))$.

In Figure 3.6, the reader can observe the process of the spread of influence when a given social influence model was selected, so both terms are related to each other, but in fact they refer to different perspectives. The distinction between these two terms is often ignored, but it is important for this dissertation. In Chapter 4, more information about spread of influence can be found, since this dissertation focuses on maximizing it.

3.7 Summary

The goal of this chapter was to present three important social phenomena and theories: diffusion of information, diffusion of innovations and social influence. Despite that they may look similar, each of them refers to different aspects of information, innovation or influence transmission over society. The main difference between them is the perspective or point of view. Diffusion of information is the most general approach and it does not consider the individual factors as strongly as latter two theories. Indeed, models for studying the diffusion of information ignore the personal factors or at least minimize them to some extent. In contrast, the diffusion of innovations and social influence focus on individuals while looking for the answer to why particular ideas, opinions, innovations or attitudes spread over society. However, what really differentiates them is the way people become convinced or influenced. In the former, diffusion of innovations, is mostly explained as a combination of individual aspects of a person that commits to some idea. In the latter - social influence - it is mainly the role of the influencer that persuades a person towards something and this pressure may be subtle or not. The rest of this dissertation will focus on social influence.

In next chapter, an important problem of social influence will be discussed. The research question that is addressed there and has been studied for more than

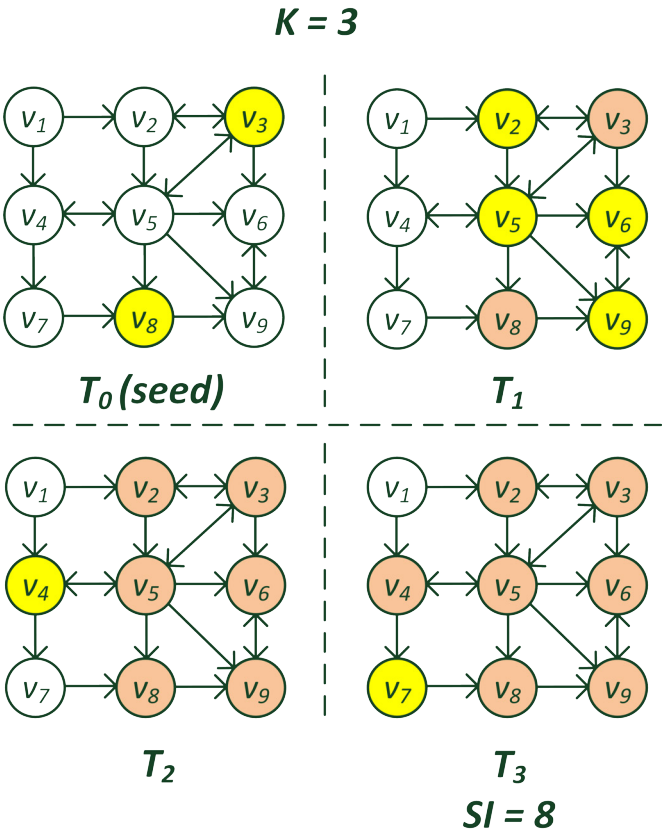


Figure 3.6: An exemplary social influence process following the linear threshold model in a social network SN. The threshold value is fixed for all nodes, the weights are equal, $\theta_v = 0.33$. At the beginning $\Phi(t_0) = \{v_3, v_8\}$, at the end of the process $\Phi(t_3) = V \setminus \{v_1\}$ where V denotes the set of vertices of SN, v_1 cannot be influenced for this combination of parameters for the LT model.

a decade is how to maximize the outcome of the influence process, the spread of influence. This chapter covers the definitions of the research problem for both cases: time-aggregated and temporal social networks, as well as the state-of-the-art in this area.

Chapter 4

Maximizing the Spread of Influence

4.1 Introduction

Different aspects of information diffusion and social influence in social networks were presented in the previous chapter. Apart from the sociological and psychological background of these processes, a number of models were presented, which are a formal representation of social influence in networks. As it was already stated, these models attempt to incorporate sociological and psychological background and with some real world data experiments it has been shown that these models can provide accurate results (Aral and Walker [2012]).

Since one can consider the social influence process as the way of convincing people to some opinions, behaviour or attitudes, one of the most interesting research problems arises - how to maximize the social influence. This problem may be considered at different levels - individual, group or society, depending on the goal. At an individual level, it could be stated as how to minimize the time required to influence a person or to reduce a cost of influencing it. In such sense, it is assumed that the influencer wants to convince a particular person, so it focuses on spending the least amount of resources to make him or her committed to the idea and the problem involves mainly psychological techniques. At a different level, when looking from the perspective of a group or the whole social

network, the problem may be stated as: who should be influenced initially ($\Phi(t_0)$) to maximize the overall spread of the idea among this group or network. In this statement of the problem it is assumed that the general process operator may freely choose nodes which will then be influenced, and by picking "proper" nodes the overall spread will be maximized. Here, often the psychological background is less important than the algorithm of choosing nodes. It means that background is being incorporated in the model of the influence and its parameters. However, the problem of influence in networks may be stated in different ways and the following chapter aims to present those variants of it as well as the state-of-the art in this area. In this chapter the problem of maximizing social influence in temporal social networks is explicitly stated, since this is the main research goal of this dissertation.

This chapter is organized as follows: The next section presents different challenges related to the influence process in social networks. In Section 4.3, the actual research problem considered in this thesis is presented and the problem of what the development of the solutions looks like is discussed in Section 4.4. The following section shows how far the recent state-of-the-art is from reaching its goal in solving this problem. Next, Section 4.5 enumerates open questions in the research domain, while the last section summarizes the whole chapter.

4.2 Social Influence Challenges

The most typical problem considered in the area of social influence in social networks is the problem of which nodes to choose to maximize the spread of influence under a given budget c (Richardson and Domingos [2002]). This topic will be explored in detail in Section 4.3. However, as presented in the previous section, different scenarios for the research on influence in social networks may also be considered. Apart from the problem of how influence an individual successfully, which was considered in Section 3.4, there are three main research questions that, unfortunately, cannot be solved at once.

In the work by Goyal et al. [2013], these challenges are presented as a constrained optimization problem summarized in Figure 4.1. The dimensions that can be optimized are the budget, the time and the number of influenced. Here,

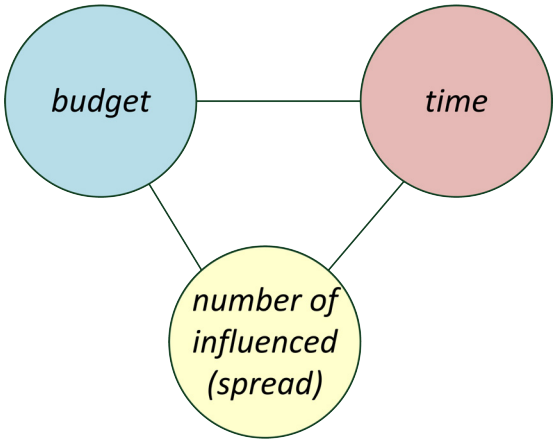


Figure 4.1: The optimization problem for the influence in social networks.

only one or two dimensions may be constrained and the third is optimized.

Below, each of the dimensions is briefly described to provide an understanding of how these dimensions are understood by most of the researchers.

Budget

The *budget* in the spread of influence problem in social networks is considered as an amount of the resource that can be spent on influencing nodes (please note that some literature prefers the more general term *activating*, e.g. see Kempe et al. [2003]). This resource is most often expressed as a budget k, such as money, gifts, conversations. However, each successful influence of a node in the network reduces the budget. Due to naming convention, in this dissertation the symbol for budget is c.

Typically, it is assumed that the amount of a budget taken for influencing a node is equal for all of the nodes in network. Sometimes it may be true, e.g. if in a marketing campaign the same product is being sent to different customers the cost of distributing the product among them is considered to be equal. On the other hand, as the influence process is a subjective one, even by spending some amount of a budget on a user, he or she may not become influenced and this susceptibility may differ from user to user. As it was already presented, those psychological aspects are included in the spread of influence model properties (such as θ for LT or $p_{v,w}$ for IC models, see Section 3.4.3 for details), but these models do not consider the

varying cost of influencing individuals in the set of seeds $\Phi(t_0)$. However, for the problem stated in the thesis, as well as in the research in this area, it is assumed that the cost of initially influencing a node is equal for all the nodes.

Time

The time constraint expressed here means that we want to influence nodes in a given time and to evaluate the results of different methods this particular time is considered as a stop condition. Typically, the models of influence described in Section 3.4.3 work until no more nodes can be influenced. This is a natural stop condition and it is reasonable for static networks and described models. Of course, when the time constraint appears, the process may be evaluated sooner. But when considering temporal social networks, the use of this hard stop condition may become more complicated, since if the network changes, the influence process may be infinite, e.g. if with every iteration new nodes join the network it could be hard to say at which moment the algorithms should stop. This is why the time constraint may be crucial for temporal social networks, as it introduces a moment in which to compare methods.

From the marketer's perspective, if they spend some budget on influencing nodes, they want to get the return from this investment in a given time, e.g. they want to have people interested in buying the product when it is offered and not discontinued (Chen et al. [2012]). On the other hand, other examples may be not so focused on the time dimension. For instance, spreading good manners among society is one of the examples. Naturally, the sooner the habits improve the better, but the time aspect is not so important here compared to the whole success of the campaign. This is why the use of time constraint is sometimes desired, but sometimes this dimension is left unconstrained. However, the reader should have in mind that for temporal social networks the use of time constraint is somehow natural, since the stop condition that no new nodes will become influenced may be wrong.

Number of influenced

The last, but most often considered dimension is the *number of influenced* (spread, see Definitions 5 and 6). The number of influenced means how many nodes were influenced or activated in the process. From the historical perspective, this dimension was the one that was maximized while the other two were left constrained or unconstrained, but Goyal et al. [2013] started to consider some other variations of the problem, e.g. by trying to minimize the time of influencing a given number of nodes.

This short introduction to dimensions should give the reader the impression that the problem of maximizing the overall spread of influence in social networks may be not the only challenge here. For instance, for some cases influencing 10% of the network minimizing the time for it or verifying whether is it possible to influence the whole network in a given time or budget, can be more important. As the next Section shows, in this dissertation the considered problem constrained the time and budget while maximizing the number of influenced nodes. Here, the constraint for time is rather for evaluation purposes than for any other reason, as was already described earlier in this section.

4.3 The Research Problem

4.3.1 Maximizing Spread of Influence

The main problem discussed in this dissertation is related to the one originating from Richardson and Domingos [2002]. The authors of this work considered the problem of finding the potential customers to be targeted by a marketing campaign that will maximize the outcome of the campaign. However, in contrast to the typical approach used in marketing, i.e. looking just at the possible sale to those people, researchers discussed the *network value* of them. The network value means that instead of focusing just on the potential targeted person, it was observed how the influenced customer will spread enthusiasm for the product to others convincing them to buy it as well. Naturally, the spread of information takes place by social network links here, so at the beginning, the direct neighbours may become influenced, then the second order neighbours and so on. In

fact, nowadays selling techniques are often based on viral marketing, so it seems that the business strongly believes in the market potential of such an approach (Kajdanowicz et al. [2013, 2014]).

In the above mentioned work, the researchers were considering the problem of building the optimal marketing campaign defined as follows. Considering a set of n potential customers, a product described by its attributes $Y = \{Y_1, \ldots, Y_m\}$, a Boolean variable indicating whether a node i will buy the product and $N_i = \{X_{i,1}, X_{i,n_i}\}$ being direct neighbours of node i, the target was to find the set of optimal actions targeting the nodes in the network that will maximize the profit understood as a number of influenced nodes. This set was referred to as a marketing plan $M = \{M_1, \ldots, M_n\}$. Since authors used the perspective of marketing campaigns, they considered the cost of the marketing to customers and the revenue from selling the product with or without marketing activities. The authors modelled the problem by means of Markov random fields and provided heuristics for choosing the customers to target. In particular, the marketing objective function to maximize is the global expected lift in profit, that is, intuitively, the difference between the expected profit obtained by applying a marketing strategy and the expected profit obtained using no strategy at all (Bonchi [2011]).

A more general question which was more often repeated in further research was stated in Kempe et al. [2003]. In this work the authors defined the influence of a set of nodes A, denoted as $\sigma(A)$, to be the expected number of active nodes at the end of the process, given that A is this initial active set A_0. The influence maximization problem asks, for a parameter c, to find a c-node set of maximum influence. From this dissertation's perspective this statement of the problem is more universal than the previous one and it was the reason why it is considered as a most popular definition of *influence maximization problem*. In the previous work, the authors focused on a detailed case of products which were defined by attributes included in the model. Here, in contrast, it is assumed that the challenge is to find an optimal set of nodes under any model of influence, which also covers the case of Richardson and Domingos [2002]. The problem here was also defined as a discrete optimization one and the formal definition of it is as follows. Given a directed and edge-weighted social graph $G = (V, E, P^m)$, a propagation model m, and a number $k \leqslant |V|$, find a set $A \subseteq V|$, $|A| = k$, such that $\sigma_m(A)$ is

maximum. In this definition P^m are the parameters related to the model, such as θ for the LT model or p for the IC model (see Section 3.4.3 for details). Unfortunately, in the same work the authors prove that the problem is NP-hard, which justifies the need to find some suboptimal solutions.

As the models considered by the authors of Kempe et al. [2003] were LT and IC, by the definition of these models, it is concluded that the only constrained dimension was the budget and the maximized dimension was the number of influenced. The time dimension did not matter here, since the models ran the process until no more influences were possible (see Section 4.2).

After these works were published in 2001-2003, a vast amount of research was conducted to provide suboptimal, efficient and scalable solutions for this problem. They will be presented in chronological order showing how the approach to tackle the problem has changed over time, see Section 4.4. However, as it will be shown, this research was applied mostly to the static social networks. Unfortunately, as the time-aggregated view of the network is used, from an information or influence propagation point of view, the most important aspect is missing - the order of contacts. As it was stated in Pfitzner et al. [2013], in these networks one assumes transitive paths, which of course do not hold in temporal or the most granular representation of networks. Moreover, as the contacts within social networks are often *bursty* (Barabási [2010]), the static representation of networks will also ignore this fact leading to wrong conclusions about the dynamic processes taking place in it. This is especially crucial while modelling the spread of epidemics, since the accuracy of predictions may strongly influence the potential actions in healthcare (Masuda and Holme [2013]). To not to loose the temporal information, researchers more and more use temporal representation of networks (Holme and Saramäki [2012]). So when looking from the perspective of social influence, the process has enough psychological and sociological complexity itself. It is also modelled by using simplified assumptions about how people become influenced (Hedström and Bearman [2009]) and using time-aggregated networks to model human interactions definitely does not help to understand the speed and direction of influence.

This is why there is a need to redefine this influence maximization problem in the case of temporal social networks and this will be done in Section 4.3.2.

4.3.2 Spread of Influence in Temporal Social Networks

The problem definition of maximizing the spread of influence in temporal social networks is conceptually similar to the one used for static networks, however the time dimension requires the formal restatement of the problem. Since this thesis considers the Linear Threshold model (see Section 3.4.3 for details), the problem definition is presented for temporal social networks and the influence model LT. Naturally, the generalization of the problem is relatively simple, as it would require only to replace the propagation model.

Let us consider spread of influence within the framework of Temporal Social Networks. A temporal social network as defined in this thesis consists of time-ordered network $TSN^K = (SN_1, \ldots, SN_p, \ldots, SN_K)$, $K \in \mathbb{N}_+$, where SNs' nodes and edges correspond to nodes' social common activities in a given time interval out of the set of intervals (see Definition 4 for details). Due to the fact that there might appear multiple events (common social activities) between v_i and v_j within a single time window T_p, the relationship is obtained by aggregation of these events and calculating the weights over edges (see Section 2.3.1). Therefore, each network SN_p is the time-aggregated one. To get more detail on what the framework for creating TSN looks like, and to read some more about them, please see Chapter 2.

Each social network $SN_p = (V_p, E_p)$, $1 \leqslant p \leqslant K$, is composed of a set of nodes $V_p = \{v_1, \cdots, v_n\}$ and a set of directed edges E_p representing relations between individuals in time window T_p. Let $N_{i_{T_p}}$ be the set of directly neighbouring and potentially influencing individuals $v_j \in V_p$ for node $v_i \in V_p$, i.e. nodes with relation to node v_i in time window T_p: $N_{i_{T_p}} = \{v_j : (v_j, v_i) \in E_p\}$. In other words, the set $N_{i_{T_p}}$ is composed of individuals who can potentially influence node v_i in time window T_p.

It is assumed that before the spread of influence begins, i.e. before the time window $T_p = 1$, a subset of individuals $\Phi(0) \subset V_0$ is selected as the seed for the influence spread. By V_0 there are denoted nodes that had been observed in the network before an influence spread was considered.

The set $\Phi(0)$ should represent a group of individuals that have already been influenced as well as the set of promoters who have certain social, economic and/or

political abilities to influence others.

It is assumed that the initial seed set $\Phi(0)$ adopts all influence before the observed spread starts. At the following time window T_1, and according to the LT model an individual $v_i \in V_1 \setminus \Phi(0)$ becomes influenced in T_1 if at least $\theta_{v_i} \in (0, 1]$ weighted sum of its neighbours' influences are in the seed set, i.e. in the set of already influenced nodes (see Section 3.4.3.2 for details of LT model).

$$\frac{\left|\Phi(0) \cap N_{i_{T_1}}(SN_1)\right|}{\left|N_{i_{T_1}}(SN_1)\right|} \geqslant \theta_{v_i} \Leftrightarrow v_i \in \Phi(1) \tag{4.1}$$

It means that set $\Phi(1)$ consists of all nodes who have been exposed to the influence, are persuaded by their neighbours and adopted the influence in period T_1.

In general, for a given period $p \in \mathbb{N}$, a not-yet-influenced node $v_i \in V_p \setminus \Phi(p)$ will be influenced in the pth window (t_p, t'_p), if

$$\frac{\left|\{\Phi(p-1)\} \cap N_{i_{T_1}}(SN_p)\right|}{\left|N_{i_{T_1}}(SN_p)\right|} \geqslant \theta_{v_i} \Leftrightarrow v_i \in \Phi(p) \tag{4.2}$$

Finally, the list of nodes influenced in the following periods is obtained: $\Phi(0), \Phi(1), \ldots, \Phi(p), \ldots, \Phi(K)$. Please note that $\Phi(0) \subseteq \Phi(1) \ldots \subseteq \Phi(p) \ldots \subseteq \Phi(K)$. Moreover, even if the condition 4.2 is not met any more, the already influenced node v_i will be still influenced, since this is a progressive process, as described in Section 3.4.3.1.

According to Equation 4.2, the final set of influenced nodes $\Phi(K)$ depends on two crucial factors: initial seed set $\Phi(0)$ and the dynamics expressed as consecutive social networks SN_p for $1 \leqslant p \leqslant K$ that determine the influencing neighbourhoods $N_{i_{T_p}}(SN_p)$ for each node v_i in the consecutive periods.

This process is presented in Figure 4.2 (see Michalski et al. [2014]). Here, information from the past is used to generate seed set $\Phi(0)$ for three types of temporal social networks: TSN^{10} consisting of ten time windows, TSN^5 consisting of five time windows and TSN^1 – a time-aggregated social network. In t_0, the initial nodes are influenced and the outcome of the spread of influence is observed in the right-hand side of this figure.

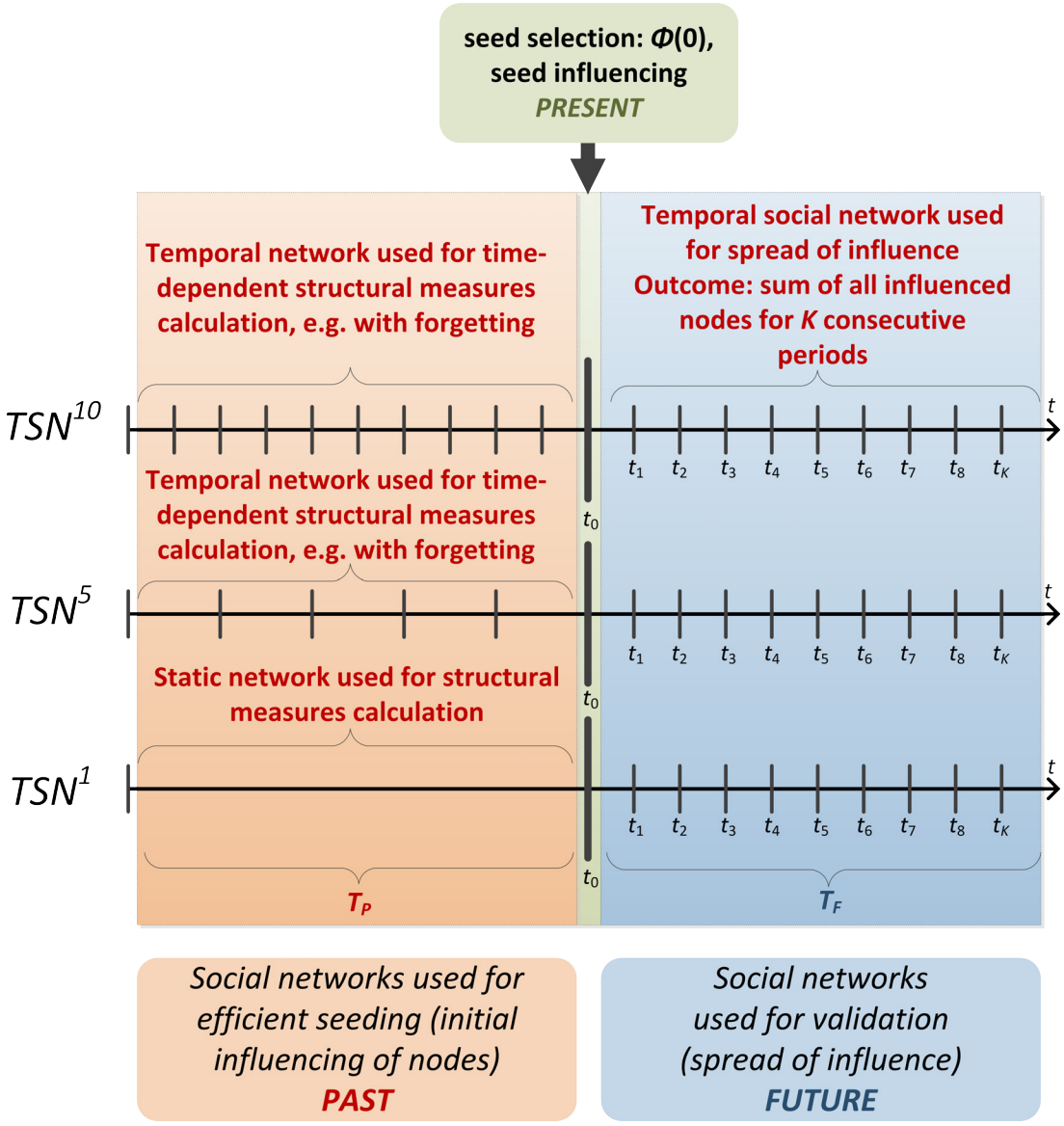


Figure 4.2: Maximizing the spread of influence in temporal social networks.

Regardless the propagation model used for spread of influence, the seed selection strategy determines the final number of influenced nodes in the network. Given the temporal social network TSN^K that consists of K social networks, the goal is to select the initial seed set of nodes $\Phi(0)$ of size c ($|\Phi(0)| = c$) in order to maximize the final number of influenced nodes $\Phi(K)$ after the Kth period in the influence propagation using the LT algorithm, see Equation 4.3.

$$\arg\max_{\Phi(0)} |\Phi(K)| \tag{4.3}$$

Due to the fact that the final set of influenced nodes $\Phi(K)$ depends on initial seed set $\Phi(0)$ and the social networks SN_p for $1 \leqslant p \leqslant K$, its exact estimation is highly complex.

The real setting of the influence spread problem might be much more complicated. In real world applications, it can be expected that seeds will be selected using historical knowledge about nodes' activity. In this case information about past activity of nodes may become an advantage, because, for instance, recently inactive nodes may be omitted from the seeds set improving general results.

Still, the dynamics of the complex networks introduce completely new problems in comparison to the static approach. In particular, we would need to address the following problems:

- As the activity of nodes may differ, even highly active nodes may become inactive just after the moment of seed selection. If a node is selected as a seed, it is expected that it will also be active later on, which may not necessarily be true, leading to wasting the marketing campaign budget.

- After the initial influencing moment, the increasing dynamics of the network in terms of new nodes appearing, may minimize the expected influence of old nodes selected as seeds.

- Due to the fact that network dynamics also include changes in edges, it may happen that these may be either helpful or harmful, i.e. influential nodes meet not previously expected non-influenced nodes, but it may also lead to undesirable outcomes - the expected node behaviour (high susceptibility to

influences or high ability to influence others) may not necessarily be valid any more.

In fact, the above mentioned problems could be solved, if the link prediction solutions (Liben-Nowell and Kleinberg [2007]; Michalski et al. [2012c]) would foresee new links with acceptable level of accuracy. This, in turn, could enable development of completely new seeding methods for temporal networks, yet, there is a lot of research still to be performed before it becomes true.

4.4 The State of the Art

4.4.1 Introduction

Below the most important works in the area of influence maximization in social networks are presented showing how the solutions evolved in time. As the problem is NP-hard (Kempe et al. [2003]), researchers focused on building heuristics that provided sufficient accuracy, efficiency and scalability. The readers interested in more detail on the nuances of the presented algorithms are referenced to Bonchi [2011].

4.4.2 Original Problem Statement

As it was stated in Section 4.3.1, the influence maximization problem was originally introduced in Domingos and Richardson [2001]. The authors posed a question on how to pick nodes and influence them with the idea of maximizing the overall spread of this idea across the network. In this work the example of a marketing campaign was used and the authors considered the network value of a customer, i.e. the benefits for the company if this customer would influence their neighbours. The influence of nodes on each other was modelled as a Markov random field (Kindermann et al. [1980]) and the obtained results revealed that this direction may be promising. In the next work on this topic authors used a linear model where the solution for influence maximization was based on solving linear equations (Richardson and Domingos [2002]). However, what this model lacked, was the iterativeness, since it reflected the joint distribution over all nodes.

Compared with the psychological research on social influence or diffusion of innovation, the process is rather iterative, so the models representing it are most often of this kind.

4.4.3 The Greedy Algorithm

The work that showed a different approach to the one presented by Domingos and Richardson was Kempe et al. [2003]. The authors assumed that the influence is more an iterative process, so they analysed two models of this kind, namely LT and IC. Making these the basis of their research, firstly they considered the hardness of the influence maximization problem and in both cases it was proved to be NP-hard. Then, by taking the advantage of the properties of *submodularity* (Schrijver [2003]) and the research on the greedy hill-climbing algorithm they showed that the greedy algorithm may outperform classic approaches based on network measures such as top degree or top betweenness. In fact, the authors showed that the outcomes of this approach may not be worse than 63% of an optimal solution $(1 - \frac{1}{e} - \epsilon)$. The greedy algorithm pseudo-code is presented as Algorithm 3 (based on Chen et al. [2009]). Here, given a social network $G = (V, E)$ consisting of vertices and edges, an initial seed set of c nodes is selected iteratively which maximizes influence. In each step of the algorithm, a single vertex is selected, such that the influence of the set $\Phi(t_0)$ and this vertex together is the greatest. Unfortunately, this algorithm has a few drawbacks. One is low efficiency – the influence is estimated with R simulation steps (Chen et al. [2009]). The other is that the algorithm is trying to pick nodes that maximize the influence in each iteration. When comparing it to the chess game, it always chooses the move that affords the best position at that moment, without considering the next move. Sometimes it may be better to look at a combination of moves or nodes rather than at a single next move to maximize the overall result, and the greedy algorithm avoids it by its nature, since it finds the local optimum. However it still provides acceptable results compared with the optimal solution, but sacrificing efficiency.

To overcome the drawbacks mentioned above, the research went in two directions. Firstly, a number of techniques were proposed to optimize the greedy

Algorithm 3 Greedy algorithm for maximizing the influence

Require: budget c, social network $SN = (V, E)$
1: initialize $\Phi(t_0) = \emptyset$
2: **for** $i = 1$ to c **do**
3: $\quad \Phi(t_0) = \Phi(t_0) \cup \{\arg\max_{v \in V \setminus \Phi(t_0)} SI(\Phi(t_0) \cup \{v\})\}$
4: $\quad$ i = i + 1
5: **end for**
6: output $\Phi(t_0)$

algorithm. Secondly, researchers started to search for new ways of maximizing the spread of influence. Below, how the greedy algorithm was improved is presented, and later on, the new ideas on maximizing influence in social networks are presented.

4.4.4 Greedy Algorithm Optimization

The work of Leskovec et al. [2007] shows there is also one more drawback of a greedy algorithm. For now it has been assumed that the cost of acquiring a single node is equal to all others, but in social networks it may not be the case. For instance, as there is an ongoing marketing campaign and their designers like to provide influential social network users with some incentives, their expectations may differ from one to another. On the other hand, there are some scenarios in which in the campaign the same products are sent to different users assuming that they become influenced, so the potential cost is equal. However, since the influence is rather a subjective than an objective process, the same gift may not result with the same user satisfaction. In contrast, there are some cases of social networks where equal cost is possible. For instance, in epidemiology it is often assumed that the cost or probability of infecting a person is the same for the whole population, but this is not the case with influence. In sensor or computer network, the cost may be equal, but the assumption of the same cost for different individuals seems to be limited.

In the paper Leskovec et al. [2007] the case of different costs for influencing each node was analysed, and it was proven that with this assumption the greedy algorithm performs badly. To overcome this limitation the authors introduced

a novel approach, *Cost-Effective Forward selection* (*CEF*) that uses the greedy algorithm and the cost-sensitive method in parallel and these are compared later to find the best one. Moreover, again by using the submodular properties of the function, the authors are able to reduce the number of possible runs of evaluation of the quality of the selected node ($Inf(\Phi(t_0) \cup \{v\})$), because they make use of the fact that the marginal increase of benefits with each added node does not increase more than in the previous evaluation. They call this approach *Cost-Effective Lazy Forward selection* (*CELF*) and compared to the greedy algorithm it is up to 700 times faster, still with acceptable results of at least $\frac{1}{2}(1 - \frac{1}{e})$ of optimal solution. Comparing it with the greedy algorithm, which proposes $(1 - \frac{1}{e} - \epsilon)$, makes CELF a very good rival.

However, at least for the IC model, there was still some space for improvement, as shown in Chen et al. [2009]. The authors decided to use random graphs in order to reduce the number of runs R (see Algorithm 3). Their approach assumes that for the IC model it is possible to reduce the graph of influences to only these edges that are potentially reachable from the set $\Phi(0)$ at the ith iteration. This change allowed the gain of an additional 15-34% improvement in running efficiency by keeping the same level of quality. More interestingly, in the same work the authors proposed a new heuristic that significantly improved the influence spread while running more than six orders of magnitude faster than all greedy algorithms - *DegreeDiscountIC*. This approach is using a degree heuristic, but *discounts* the degree of a node v considered as a seed by the value of already influenced neighbours of v. Since there exists a non-zero probability that this node will become influenced by one of its influenced neighbours, it makes it less attractive as a seed.

Goyal et al. has shown that further exploitation of submodularity may lead to even better results for greedy algorithms. In Goyal et al. [2011b] an extension of CELF algorithm was presented; it leads to at least 35% gain in performance. The idea is to store a heap for all non-selected nodes that contains information not only about a marginal gain of a particular node, but also the marginal gain of the best node from those evaluated before this node. Due to this trick there is no need to recalculate the marginal gain of a node if this node was not selected, resulting in less iterations of the algorithm.

However, if analysing advances in science, the problem of maximization of the influence is the most popular one. The next section describes the approaches in this area that do not improve the greedy techniques but propose something completely new.

4.4.5 Avoiding Greedy Search

One of the approaches to avoid greedy search has already been mentioned, it was the *DegreeDiscountIC* algorithm (Chen et al. [2009]). The same authors proposed another method (Chen et al. [2010]), *Maximum Influence Arborescence* (MIA), which also exploits submodularity. However, in the case of the IC model, for each pair of nodes, maximum influence paths are calculated and then the authors discard these path below the specified threshold of influence, focusing only on local regions of influence. Then the authors join these paths, creating tree structures which do need to be updated often and the calculation of influence spread may be done recursively. This leads to significant gains in the effectiveness of the algorithms compared to others and introduces no loss in terms of quality. Moreover, the threshold parameter may be interpreted as a way of controlling the time of influence and the overall spread.

Another work (Jiang et al. [2011]) also avoids the greedy approach while outperforming it in terms of speed (2 to 3 orders of magnitude) and bringing similar results in accuracy. The authors claim that this is the first approach that uses simulated annealing (Metropolis et al. [1953]) in solving the problem. In the case of influence maximization this approach starts with random seeds and then tries to move into the space of possible solutions (initial seeds) towards the local minimum by swapping, at most, one node in the seed set until the stop condition is applied.

An interesting insight into the problem is given by Shakarian and Paulo [2012] where the authors propose an algorithm that guarantees to activate (influence) the whole network. The solution does not find the minimal set of seeds, but its outcomes may be compared to the budget c - if the seed set is less or equal c, the algorithm will fulfil the requirements and, moreover, the whole network will be influenced. The approach is based on removing edges in the graph (using the

idea of shell decomposition, see Carmi et al. [2007]), but it also guarantees to influence the whole network.

The last mentioned algorithm in this section is Simpath which is intended to maximize the influence under the LT model (Goyal et al. [2011c]). This algorithm operates on paths of influence in network by assuming that most of the influence is local. Results reveal that the algorithm outperformed the MIA method, considered as the state-of-the-art in the task of influence maximization for static networks (Chen et al. [2010]). Another approach is presented by Liu et al. [2013] where the authors focus on measuring independent influence in contrast to the strategies presented above, i.e. they focus on the individual's chances to influence others, i.e. they treat individuals' influence separately.

4.4.6 Towards Temporal Social Networks

From the author's perspective, all the techniques presented above may be considered as purely structural ones. Here, the researchers do not use any kind of attributes of nodes other than their structural properties, such as location in the network or interconnectivity with other influenced nodes. The only parameter that may differentiate the nodes is the cost of influence, but in most cases it was assumed to be uniformly distributed. This approach of using just the network's structural properties makes the proposed algorithms universal, since they do not require any network-specific attributes. As it will be shown later, there is also another emerging direction in this research that is using the messages' content, so the merit of the social network communication. It is worth emphasising that it is worth noticing how many of the presented approaches took advantage of the submodularity property.

However, all algorithms presented above suffer from one drawback which makes them just rough simplifications of reality. Here, the network dynamics are not considered, and as it was already stated in Section 4.3.1, ignorance of this fact may lead to wrong conclusions about the outcomes of the process. So later in this section it is shown how the work on influence maximization was developed for the dynamic scenario. Since most of the research presented below makes the approaches more data-dependent, the reader must bear in mind that

in contrast to the purely structural algorithms described above, the application of the approaches introduced below may be limited, since the researchers have not always had full information about social networks (e.g. communication content). Of course this is not an argument to avoid this direction, but just a loose remark.

4.4.7 Learning Influence Probabilities

Before thinking of maximizing the spread in real world temporal social networks, there should be at least one limitation overcome, which is how to assign the influence probabilities making them more aligned to reality. As it was presented in Section 3.4.3, the influence probability is often drawn from a distribution or hard-wired. When thinking of real world scenarios, this assumption makes the results of modelling the spread of influence questionable. In this area there are just a few papers which attempt to deal with this problem.

The research on this topic started with the work of Tang et al. [2009]. Here, the authors try to avoid learning influence probabilities from the network position of a node, since they assume that different peers of a node may have a different influence on it. For instance, our friends may be more influential in the area of private life (trends, friends etc.), while relatives from work may have a stronger influence in company-related areas. Taking this into account, the authors decided to analyse the content of the communication to build a model of *Topical Affinity Propagation* (TAP). This approach attempts to assign influence values over edges between nodes which are topic-specific. So in this case a node may have multiple edges with its neighbour and each of them represents a different topic altogether with different influence weights. The authors use *factor graphs* (Kschischang et al. [2001]), in which the observation data are cohesive on both local attributes and relationships. Moreover, to make the approach scalable, they do the following: they define a *Topical Factor Graph* (TFG), then they introduce *Topical Affinity Propagation* and finally they try to make the approach scalable for large networks, either by using Map-Reduce approach or a parallel update rule. The main goal of the proposed idea is expert identification, but this approach is suited well to learning real influence probabilities, as the real-datasets experiments show.

Another approach in this area was proposed in Saito et al. [2008]. Here, the

authors focus on the IC model and they calculate the likelihood of so-called *episodes*, which are in fact nodes that became influenced in consecutive time-windows. Then they compare neighbouring episodes - the one in time t, which is $D(t)$ in authors' notation, and the next one in $t+1$ (defined as $D(t+1)$) to see whether the neighbouring nodes were in $D(t)$ and $D(t+1)$. If so, it is probable that the newly influenced node v_z from $D(t+1)$ was influenced by v_y from $D(t)$ in time $t+1$. It is possible, because the IC model gives just a single chance for a node to influence its neighbours, so the influence may happen only shortly after the node itself becomes influenced. Then the authors use the expectation maximization technique to obtain the values of likelihood functions of θ. Experiments conducted on a blogging platform confirm that this approach may be right in terms of obtaining the influence probabilities by learning from past data.

The last work presented here is the work of Goyal et al. [2010]. In contrast to the previous work the authors generalize their approach to every influence model following the submodularity property, which makes it more universal (e.g. covering LT and IC). As the reader may remember, submodularity was the property which allowed the introduction of many improvements in the area of maximizing the spread of influence, see Section 4.4.4. In this work the authors combine two sources: the temporal social network and an action log which represents the activity of users. In detail, the action log is defined as a relation containing tuples (u, a, t_u), where u represents a node $\in V$, a - an action from *the universe of actions* and t_u the time when the user u performed the action a. By proposing models for capturing static and dynamic influence the authors are able to compute the probabilities of influence and they reduce the number of scans of typically-huge action log. Moreover, they are able to predict at which time a user will take an action.

Now that we know how to learn the influence probabilities by using real world data, it is worth examining how influence maximization solutions are applicable to temporal social networks.

4.4.8 Social Influence Maximization in Temporal Networks

The approach which combines the action log and a social graph presented by Goyal et al. [2010] was later extended to maximize the spread of influence in Goyal et al. [2011a]. Firstly, the authors show that using real
world data (action logs or history of past propagations) is crucial, since it is only by knowing real influence probabilities that any algorithm for influence maximization can be accurate. So the initial assumption is that these probabilities have to be computed by real world data. Then the authors propose the so-called *Credit Distribution model* (CD) having different assumptions comparing to the LT and IC models, since it considers actions as a source of influence in network. The authors introduce propagation graphs which include nodes that were neighbours in graph E and performed the same action but in a different time. Here, a node performing an action earlier may be considered as a potential influencer of its neighbours taking this action later. So, in fact, the initial graph is static, but the actions introduce the dynamics here. Under the credit model for each action performed by a node, all nodes that took this action earlier and are neighboured to this node receive credits for being potential influencers, and this is a recursive operation. Then the authors try to choose nodes which will maximize influence in the whole network under the so-defined model offering $(1 - \frac{1}{e})$ approximation compared to the optimal solution and the scalability as well. The biggest achievement here is the lack of need to perform costly Monte Carlo simulations, but it is because of the different model definition. However, the results show that the CD model and the method to choose seeds allow to outperform common approaches for LT and IC influence maximization offering also speed improvement. It is also worth studying this paper because the authors compare the seed sets provided by different models.

In a work of Mathioudakis et al. [2011] the authors use a past propagation log and a social graph to find c the most influential links, i.e. links that will maximize the propagation. However, what they do makes a significant reduction in the search space by benefiting from sparsification. They apply their approach to the IC model and propose a *Spine*, dynamic programming algorithm, which proposes a significant improvement in speed, offering accuracy close to optimal.

Spine is structured in two phases. During the first phase it selects a set of arcs D_0 that yields a log-likelihood larger than $-\infty$. This is done by means of a greedy approximation algorithm for the Hitting Set NP-hard problem. During the second phase, it greedily seeks a solution of maximum log-likelihood, i.e. at each step the arc that offers the largest increase in log-likelihood is added to the solution set (Bonchi [2011]).

An interesting approach to considering time-varying influence is proposed in Liu et al. [2012] where the authors consider the delayed influence process, i.e. the influence of a node on its neighbours may vary in time. It places the problem closer to reality, where people pay more attention to recent incidents than to older ones. The authors propose *Influence Spreading Paths* as a method of measuring the influence of a node, i.e. $ISP(u, S)$ represents all spreading paths that end with user u. By using them authors compute the activation probability of a user u and thanks to that they are able to find seeds faster than by using the greedy algorithm. So the time factor incorporated here is not the time reflecting the dynamics of the network but the changes in influence probabilities. However, it is another way of representing the network dynamics and as such it can be used for solving the problem in a dynamic environment.

The problem of influence maximization in temporal social networks was just recently explicitly stated in Aggarwal et al. [2012]. As the authors claim, to their knowledge they propose the first set for temporally sensitive methods for influence maximization. In this work researchers use the transmission matrix which contains the time-dependent functions for influence spread to find a solution for two separate problems. Firstly, they would like to pick c nodes at time t_1 to maximize the influence at time t_2 - this problem lies closer to the classic influence maximization problem, but it incorporates the time factor. Secondly, when observing the influence spread at time t_2 they would like to establish which nodes most probably were responsible for the influence spread at time t_1. To deal with these problems, they introduce *Backward and Forward Influence Algorithms*. When looking at the influence maximization problem, the authors try to solve it similarly to the greedy algorithm. But now, each iteration means another time-step. In conducted experiments it is shown that the time-dependent solutions outperformed the static ones showing that this direction should be further exploited.

By trying to take advantage of the network dynamics, preliminary research has been published (Michalski et al. [2014]) which confirms the results obtained by Aggarwal et al. [2012]. Here, the authors decided to split the temporal social network into multiple ordered social networks and evaluate different seeding strategies. Results show that it is possible to increase the number of influenced nodes by using higher granularity and time-dependent measures. It was the basis for further research extended in this dissertation.

It is worth to refer to two more works in this area which tackle the problem differently. In Li et al. [2012] researchers attempt to find successor nodes for removed seeds. It is a relatively different research question than in a typical influence maximization problem, since now the budget c will increase, but this approach incorporates the dynamics of networks showing that considering it is crucial. In Jankowski et al. [2013b] authors try to take into account the availability factor of nodes, which indeed is the embedded dynamics of the network, attempting to improve the overall influence spread in networks. Again, experimental results confirm that this direction helps the seeding strategies in obtaining better results.

There is also one more direction that can be studied in order to try to find the answer on this research question - control theory. When treating the temporal network like a dynamic system (Dralus and Świątek [2009]), the optimization task could be stated in this research domain. To solve this optimization task it could be then possible to benefit from the mature apparatus of the control theory (Helmke et al. [1994]; Rao and Singh [1979]). However, the major problem is that, as for now, no models that could describe the temporal networks as dynamic control theory model exist, at least to the dissertation author's best knowledge. Yet, by using this bridge to another research domain, some advances could be made.

At this point it is worth emphasizing that the idea of maximizing spread of influence in temporal social networks is a relatively new one, but as the above literature review shows, the idea of considering the dynamics of networks in the influence process is important and already some solutions are being proposed. However, from author's perspective the work on it has just begun and we should expect rapid development in this area shortly. One of the reasons is that the dynamics in social networks are something natural rather than unusual and it is

already agreed that the influence maximization problem should be considered in this real world setup.

4.4.9 Summary

The above presented state-of-the-art in the area of maximizing the spread of influence in social networks shows how various ideas evolved over time. As the greedy algorithm (despite its acceptable or even surprisingly good quality) lacked performance, many approaches were introduced to overcome its limitations. On the other hand, there are also works that try to avoid greedy search and some of them were quite successful. Still, all these solutions were applicable only to static networks. The work on seeding strategies in temporal social networks began in 2012 and up to now studies in this area may be considered as preliminary. More works are expected to appear in the coming years.

In order to try to briefly summarize how the idea of maximizing the spread of influence has developed over time, Table 4.1 has been compiled.

Table 4.1: The development of the influence maximization problem and solutions over time.

Year(s)	Description	Comments	Work(s)
2001	The statement of the influence maximization problem as a marketing campaigns challenge - measuring the network value of the customer	Bayesian modelling, linear model, building marketing plan	Domingos and Richardson [2001]
2003	Considering the spread of influence as an iterative process, the most popular statement of the influence maximization problem, showing the greedy algorithm results	Proving NP-hardness of the problem, LT and IC models used, $(1-\frac{1}{e}-\epsilon)$ approximation	Kempe et al. [2003]
2007	Considering varying costs of influencing nodes, introducing CEF and CELF methods joining the greedy algorithm and a cost-sensitive approach	Taking the advantage of submodularity of gain to optimize the speed of the greedy algorithm	Leskovec et al. [2007]
2009, 2010	By basing on the influence paths escaping from the greedy algorithm regime, Maximum Influence Arborescence method proposed	Submodularity again	Chen et al. [2009, 2010]
2011	Improving the CELF method by keeping the heap of not selected nodes to not to recalculate the gain in the next iteration	Further exploiting of the submodularity property	Goyal et al. [2011b]
2012	The first definition of an influence maximization problem for dynamic networks, preliminary research on solving the problem, comparison with static methods	Another example of taking the advantage of submodularity	Aggarwal et al. [2012]
2012, 2013	Compensatory seeding methods for static and temporal social networks, finding the nodes which can replace the missing seeds	Taking the advantage of the past activity in networks	Jankowski et al. [2013b]; Li et al. [2012]

4.5 Open Questions

The research on influence maximization in the social network may be considered as being in the early stages. There are numerous reasons why such an opinion is justified and they are briefly presented below.

4.5.1 The Role of Network Dynamics

One of the most important questions is whether the network dynamics are positive or negative for the influence spread. Naturally, the dynamics in social networks happen and nothing can be done to stop it, since this is a process that cannot be fully controlled. However, as presented in Masuda and Holme [2013], in some domains, e.g. epidemics, there is still the debate whether the dynamics amplify or slows down the spread in networks. It has already been shown that the natural effects occurring in networks, such as burstiness, affect the threshold models strongly (Takaguchi et al. [2013]), but it is still not the answer to this question.

Trying to solve the problem, two approaches may be used. Firstly, an empirical research may be conducted whereby using real networks of different kinds but with similar structural properties it could be possible to evaluate whether the dynamics of users influence the influence process. This kind of result, unfortunately, will lack generalization, so to overcome this limitation there is a strong need to evaluate this idea in artificial networks. To tackle the problem there is a need to explore the field of *dynamic network analysis* (DNA, Carley [2010]) in terms of models of network dynamics (Aggarwal and Subbian [2014]) and for given models of network evolution, evaluate different models of spread of influence.

Naturally, by saying „*positive or negative factor*" it is not necessarily true that maximizing the spread of influence always has a positive effect. In some domains, such as epidemiology, it can either be a good or bad outcome. For instance, the trend of avoiding vaccinations may be considered as non-desired, so the maximization of it may be harmful for society. Conversely, the idea of vaccinating children has a widespread positive effect. There are also works that consider whether just the process of incentivisation, i.e. initially influencing others to convince them to some idea, is either positive or negative for the process outcome (Michalski et al. [2012b]).

4.5.2 Outperforming the Greedy Algorithm

As the literature review presented in this Section shows the advances in research on maximizing the spread of influence it is observed that the barrier of the greedy algorithm is $(1 - \frac{1}{e} - \epsilon)$ of the optimal solution (see Section 4.4.3 for details). Further research in the area couldn't outperform the results of this approach in terms of quality, but dramatically reduced the computational time, mostly by benefiting from the submodular property of gain function. The next open question is whether is there any chance to introduce new methods that will provide better results than the greedy algorithm. In fact it is observed that nowadays' research in this area is moving towards temporal social networks and this direction will be exploited in the coming years. So, probably we will then see a *new beginning* in the area of spread of influence maximization, but devoted to solving the problem in the dynamic domain. Still, some methods of seed selection for static networks are also awaited, since some non-social networks are mostly static, e.g. telecom infrastructure or sensor networks,at least to some extent.

4.5.3 Network-dependent vs. Universal Methods

Another question is related to general strategies in this domain. As one can see in the literature review, there are two general methods of finding the solution. One is using only structural properties of nodes (the greedy method and its successors), while the other one takes into account the content in the social network (e.g. TAP method, see Tang et al. [2009]).

Naturally, using only structural properties the proposed methods is more universal, since it works for a given social influence model ignoring the network character and type. In this sense those algorithms have one advantage over the methods that analyse the content - universality. On the other hand they cannot benefit from the data to attempt to achieve better results. Since now the only benefit that is gained from using the content of information exchange is the time - still it is hard to outperform the greedy algorithm. Naturally, as the current algorithms can be in two to three orders of magnitudes faster than the greedy algorithm it makes them applicable for larger networks and this is a significant advance. Still, one could expect that using the content of the information may

lead to better results in terms of accuracy.

4.5.4 The Adequateness of Social Influence Models

As the models of social influence are expected to model this process, it is expected that they do it accurately for certain networks or (ideally) network models. The impression remains that there still exists a need to confirm that these models are realistic. In fact, as the process of influence is complex and not independent from other sources and processes, it is hard to perform the experiments that will show how accurately the process is modelled by particular models. Still, as presented in Section 3.4.3 few models fit the reality well in some cases. There still exists space for more work that will attempt to evaluate models of social influence and their parameters against real world datasets. Also, the problem of learning individual influence probabilities (see Section 4.4.7) requires further research.

4.5.5 Optimal Solution in Temporal Social Networks

The static case in the influence maximization problem seems to be relatively easy in terms of defining what is the optimal solution for the problem. According to the problem statement in Kempe et al. [2003] it is the set of nodes within a given budget c that maximizes the spread of influence. As the network is static, the best possible solution may be simply to let to reach all nodes in the network or all nodes that can be influenced under a given propagation model. Now, moving on to the case of temporal networks, the question arises what actually are all those nodes in the network. Since the network may grow and shrink, we only know of the past nodes which may be selected as seeds. This set of seeds should maximize the influence in the network at some future time t. So, to evaluate the proposed methods we should use:

- All nodes that we know of from the beginning of learning until time t?
- All nodes that appeared from the seeding moment until time t?
- Only nodes that are active in the last time-frame?

In this work one of the definitions of the optimal solution was used (see Section 4.3), but that is nor to say that other approaches are wrong. It is believed that this element of the problem will continue to be debated among researchers.

4.6 Summary

In this chapter it was shown that in the area of social networks one can define multiple challenges, not only that of influence maximization (Section 4.2). Then, sticking to the problem of influence maximization, the formal definition of the problem for the dynamic case was provided (Section 4.3). In Section 4.4 multiple solutions for the static case were presented and just a few for the dynamic one, since this problem is still far off being fully studied. Section 4.2 shows that there are still some open questions that definitely will spark some discussion in the upcoming research in this area. The next chapter focuses on experiments, where different temporal social networks were examined in order to uncover how the granularity influences the network properties.

Chapter 5

The Properties of Experimental Temporal Social Networks

5.1 Introduction

This chapter presents how the type of temporal social networks influences the structural properties of social networks of which TSNs consist. Initially, the role of this chapter may be considered rather descriptive, but as experiments conducted in Chapter 6 show, the variability of network measures' values may be further successfully exploited to maximize the spread of influence in the network.

This chapter is organized as follows. In Section 5.2 the general set-up of the experiments is presented, including the description of datasets used in the experiments, the configuration for Time-limited Event Sequences as well as for Temporal Social Networks. Moreover, since there is no generally available framework for performing the research on temporal social networks, the developed framework is presented in the same section. In Section 5.3 it is shown how the properties of temporal social networks change when the resolution of the network varies. Lastly, Section 5.4 summarizes the chapter. The purpose of this chapter is to present the most important aspects of examined temporal social networks which were later used to develop the method for maximizing the spread of influence in temporal social networks. This method is extensively studied in Chapter 6.

5.2 Experimental Setup

The experimental part of this dissertation aims to show whether and how it is possible to benefit from the temporal networks to maximize the spread of influence. This research is split into two parts. In this chapter, how the temporal social networks' properties behave for different configurations of time windows, and how sensitive these networks are to changing granularity, is analysed. This part may be considered a general analysis of temporal networks' properties evaluated on five different real world datasets. The next part, presented in the next chapter focuses on the problem of maximizing the spread of influence in temporal social networks. But, for the sake of clarity, the networks' configuration of those two parts is presented in this chapter.

5.2.1 Definitions and Research Methods

The definition of temporal social network used in this thesis is presented in Definition 4 and discussed in Section 2.6, but for readiness purposes it is repeated below.

A Temporal Social Network TSN^K on Event Sequence ES (see Definition 1) is a sequence of time-ordered component Social Networks SN_p, such that:

$$TSN^K = (SN_1, \ldots, SN_p, \ldots, SN_K), \quad K \in \mathbb{N}_+. \tag{5.1}$$

In this thesis a component Social Network SN_p (see Definition 2) is extracted from Time-limited Event Sequence TES_{T_p} (Definition 3), as presented in Section 5.2.3. The time order is non-descending, i.e. $\underset{1 \leqslant p < K}{\forall} t_p \leqslant t_{p+1}$.

Each graph SN_p, $1 \leqslant p \leqslant K$, is composed of a set of nodes $V_p = \{v_1, \cdots, v_n\}$ and a set of directed edges E_p representing directed relations between nodes in time window p : $E_p = \{(v_i, v_j) | v_i, v_j \in V_p\}$. An illustration of an exemplary temporal network following this definition is presented in Figure 5.1.

The network presented in Figure 5.1 consists of seven time-ordered graphs and each period (t_p, t'_p) is of equal duration. In this example those periods are non-overlapping, i.e. the period (t_p, t'_p) ends just before (t_{p+1}, t'_{p+1}) starts, so $closure_l =$ "[" and $closure_u =$ ")". So, formally, SN_1 consists of nodes from the time interval

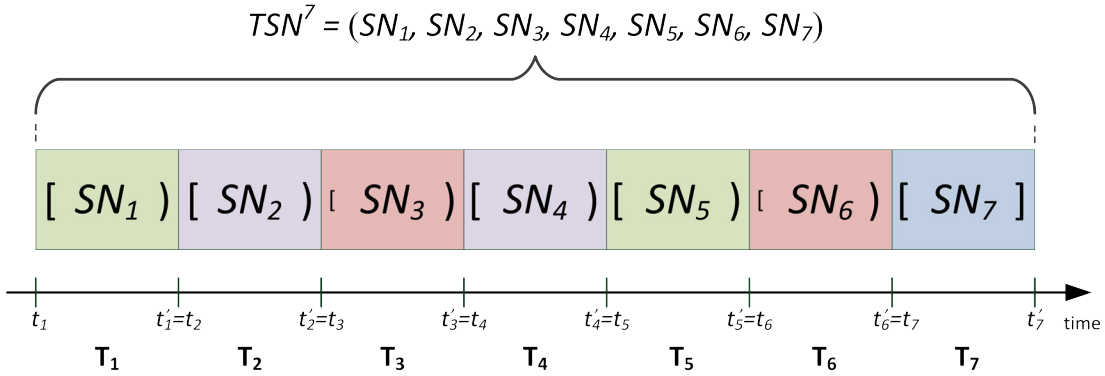


Figure 5.1: An exemplary temporal social network.

$[t_1; t'_1)$, while SN_2 consists of nodes from time interval $[t_2; t'_2)$ (the same closure types) and so on. The last network, SN_K, has a different $closure_u$ type, namely "]", in order to cover the whole event sequence. The double notation for these time intervals is helpful when a single social network is discussed as there will be no references to other time intervals' indices. However, for simplification of notation, the T_p notation for periods is also introduced.

As it has already been mentioned in Chapter 2, this is just one of the approaches used in representing temporal networks. But, as further experimental results show, this one was helpful in the task of maximizing the spread of influence. Naturally, it is not said that by using contact sequences or interval graphs (see Holme and Saramäki [2012]) the results will be worse. However, the LT model in its classic definition of Granovetter [1978] considers the static graph as an underlying layer when evaluating whether a particular node will be influenced or not, so the usage of this model was limited to temporal social networks as defined in this thesis. Just recently there was one work published that migrates the LT model to the space of temporal networks (Karimi and Holme [2012]), but when using this model of influence it is impossible to compare obtained results with other common methods of maximizing the spread of influence.

Later in this chapter temporal social networks as defined here will be generated based on a set of real world datasets. These temporal social networks will be evaluated in terms of stability of network measures and other properties to find out whether they are stable across the split types and datasets or are sensitive to

bursty activity in the network. This knowledge will then be used in Chapter 6 to maximize the spread of influence in temporal social networks.

The next section presents the datasets and their properties as well as the properties of time-aggregated networks based on these datasets to give a rough view of the characteristics of networks. Then, as TSNs have to be defined over Event Sequence or Time-limited Event Sequence, Section 5.2.3 presents how TESes were selected. Lastly, Section 5.2.4 presents how these datasets were used to construct temporal social networks.

5.2.2 Datasets Description

All of the datasets evaluated in this thesis were obtained from the repository KONECT (the Koblenz Network Collection)[1]. This repository offers variety of network datasets and some of them have timestamped edges. The experiments in this thesis were conducted using five datasets published in this repository and they are introduced and described below. Datasets are denoted as D_z, where $1 \leqslant z \leqslant 5$ and these datasets were later used for building temporal social networks as presented in Section 5.2.4.

For each dataset there is a separate section devoted, in which the description of the dataset is accompanied by the table presenting the most important properties of the dataset, and distributions of selected measures. The most important measures were described in Section 5.3.3.2. For the description of others, it is suggested to look at Wasserman and Faust [1994]. The figures in this chapter were presented here courtesy of Dr. Jérôme Kunegis - the author of the KONECT project who prepared them. At the end there is a short summary of the datasets presented in Section 5.2.2.6.

5.2.2.1 Manufacturing emails

Description: This is the internal email communication network between employees of a mid-sized manufacturing company (Michalski et al. [2011b]). The network is directed and nodes represent employees. The left node represents the sender and the right node represents the recipient. Edges between two nodes are

[1] http://konect.uni-koblenz.de

individual emails.

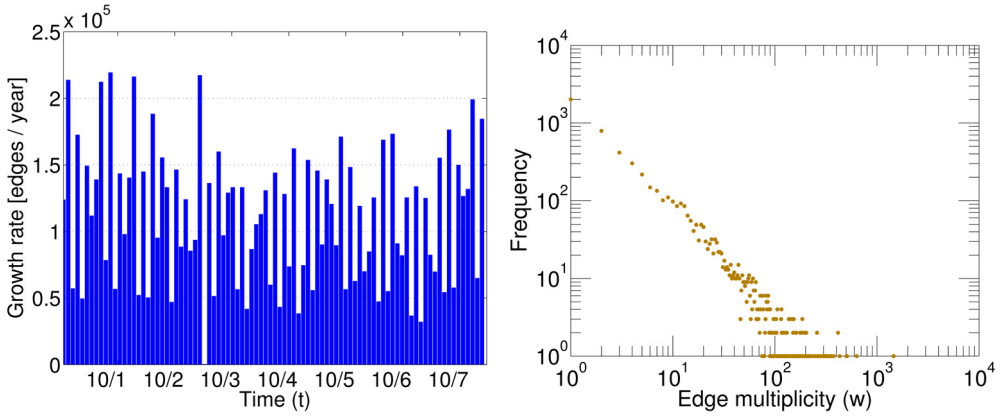
URL: http://konect.uni-koblenz.de/networks/radoslaw_email

Table 5.1: Properties of the dataset D_1 - manufacturing company emails

Property	**Value**
Period	2010-01-01 . . . 2010-09-30
Individual type	Employee
Event type	Email
Format	Directed
Edge weights	Multiple unweighted
Size	167 vertices (employees)
Volume	82,927 edges (emails)
Average degree (overall)	993.14 edges / vertex
Maximum degree	9,053 edges
Reciprocity	8.760373×10^{-1}
Largest connected component	167 vertices (network connected)
Relative largest connected component	1
Largest strongly connected component	126 vertices
Power law exponent (estimated)	4.6110 ($d_{min} = 53$)
Gini coefficient	61.9%
Clustering coefficient	5.412664×10^{-1}
Diameter	5 edges
90-percentile effective diameter	2.50 edges
Median shortest path length	2 edges
Mean shortest path length	1.96 edges
Preferential attachment exponent	1.2284 ($\epsilon = 2.4955$)

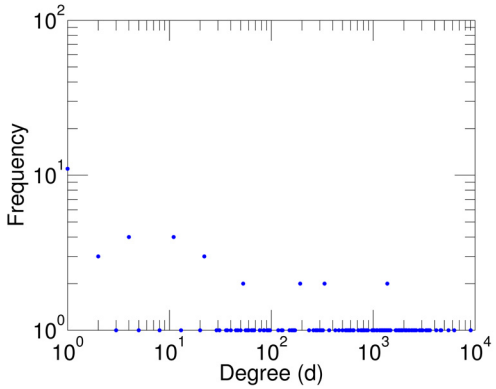
5.2.2.2 Enron email network

Description: The Enron email network consists of 1,148,072 emails sent between employees of Enron between 1998 and 2002 (Klimt and Yang [2004]). Nodes in the network are individual employees and edges are individual emails. It is possible to send an email to oneself, and thus this network contains loops.

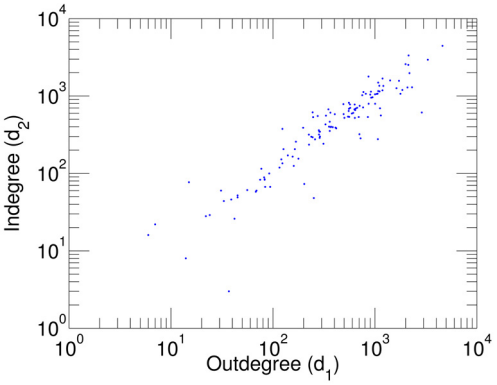


(a) *Temporal distribution*

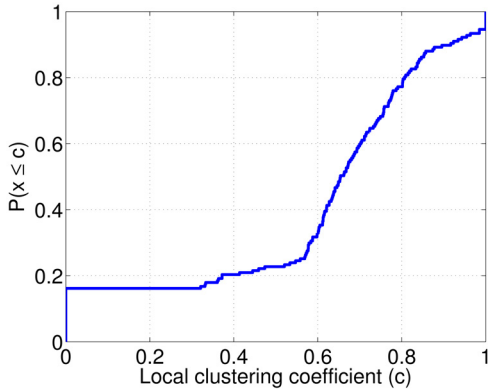
(b) *Edge multiplicity distribution*



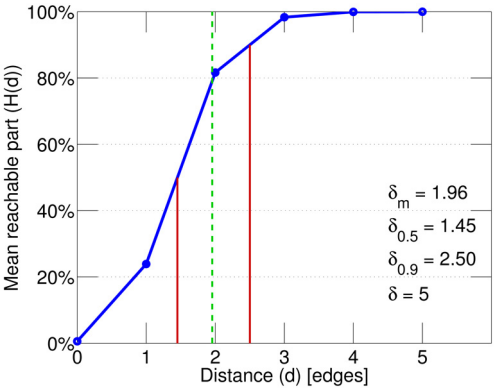
(c) *Degree distribution*



(d) *Out/in-degree comparison*



(e) *Clustering coefficient distribution*



(f) *Distance distribution*

Figure 5.2: Distributions for the dataset D_1 - manufacturing company emails.

URL: http://konect.uni-koblenz.de/networks/enron

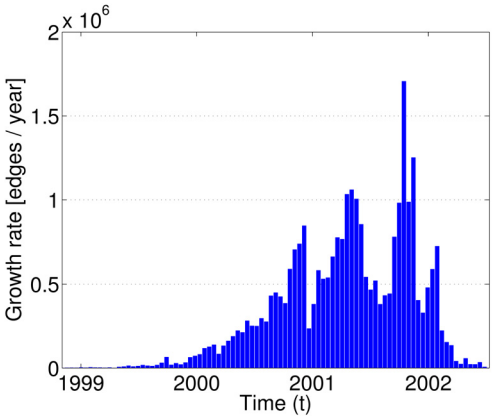
Table 5.2: Properties of the dataset D_2 - Enron email network

Property	**Value**
Period	1998-11-02 . . . 2002-07-12
Individual type	Employee
Event type	Email
Format	Directed
Edge weights	Multiple unweighted
Size	87,273 vertices (users)
Volume	1,147,126 edges (emails)
Average degree (overall)	26.629 edges / vertex
Maximum degree	38,778 edges
Reciprocity	1.46678×10^{-1}
Largest connected component	84,220 vertices
Relative largest connected component	96.7%
Largest strongly connected component	9,160 vertices
Power law exponent (estimated)	1.7610 ($d_{min} = 2$)
Gini coefficient	90.8%
Clustering coefficient	7.1948×10^{-2}
Diameter	14 edges
90-percentile effective diameter	5.80 edges
Median shortest path length	5 edges
Mean shortest path length	4.89 edges
Preferential attachment exponent	0.78589 ($\epsilon = 3.8168$)

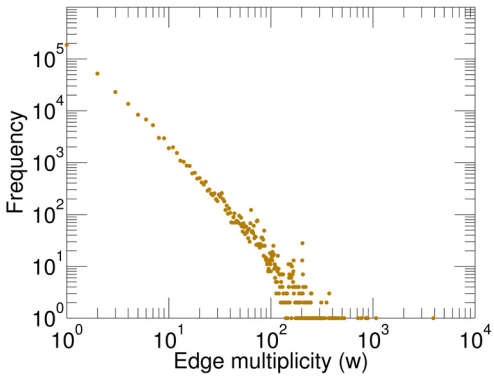
5.2.2.3 University of California messages

Description: This directed network contains sent messages between the users of an online community of students from the University of California, Irvine (Opsahl and Panzarasa [2009]). A node represents a user. A directed edge represents a sent message. Multiple edges denote multiple messages.

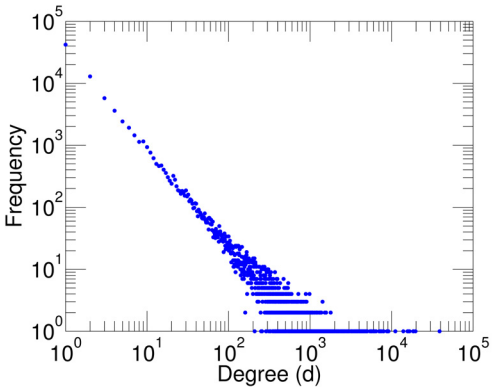
URL: http://konect.uni-koblenz.de/networks/opsahl-ucsocial



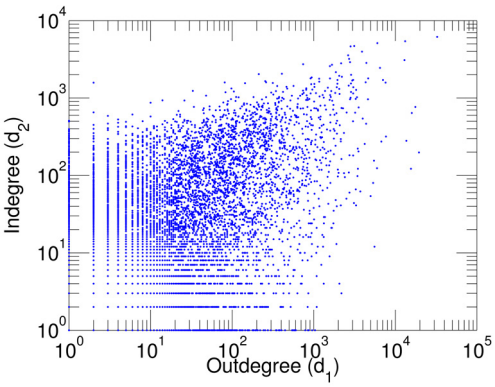
(a) *Temporal distribution*



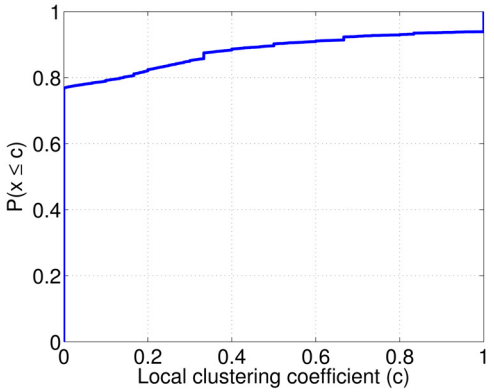
(b) *Edge multiplicity distribution*



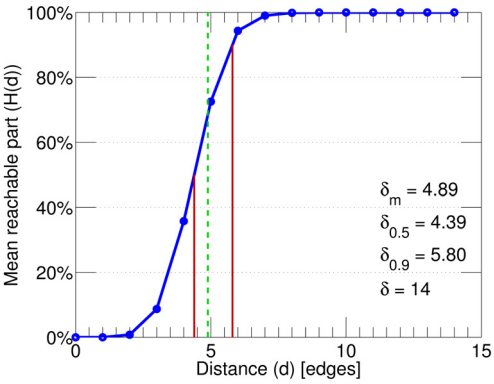
(c) *Degree distribution*



(d) *Out/in-degree comparison*

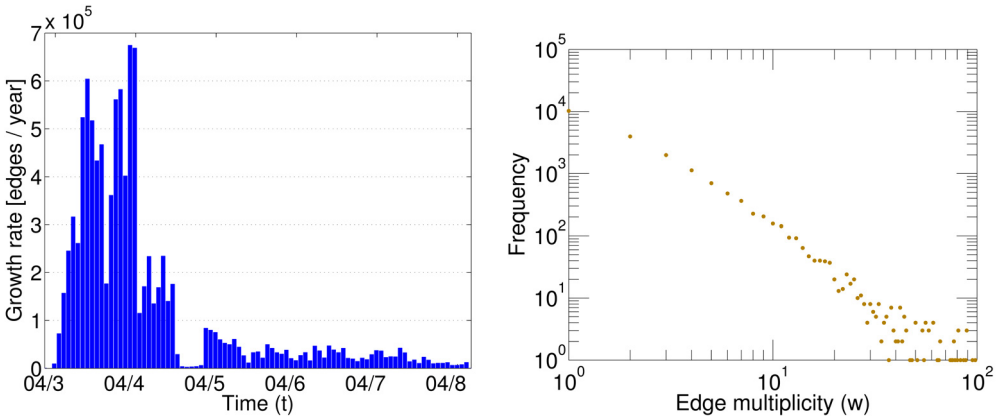


(e) *Clustering coefficient distribution*



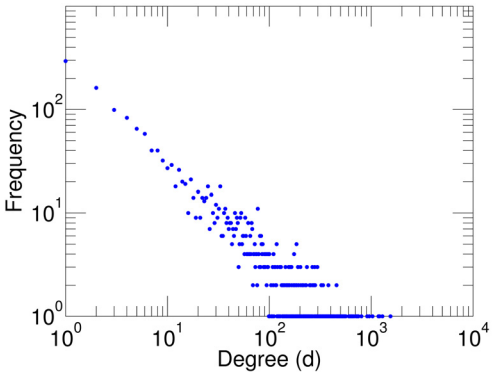
(f) *Distance distribution*

Figure 5.3: Distributions for the dataset D_2 - Enron company emails.

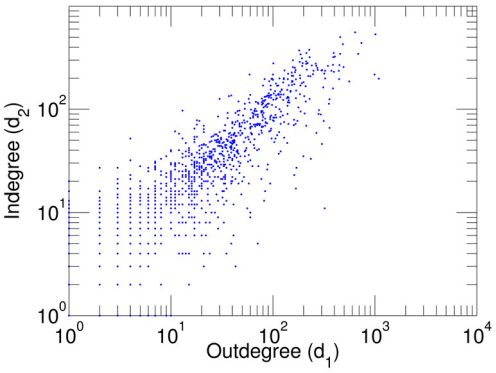


(a) *Temporal distribution*

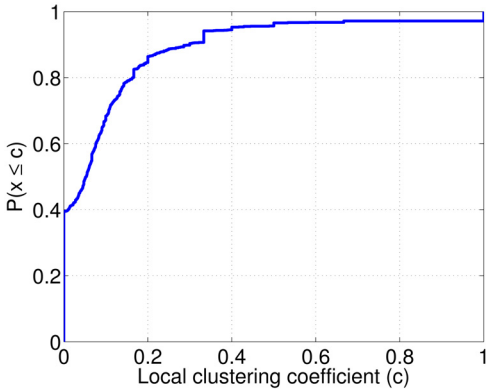
(b) *Edge multiplicity distribution*



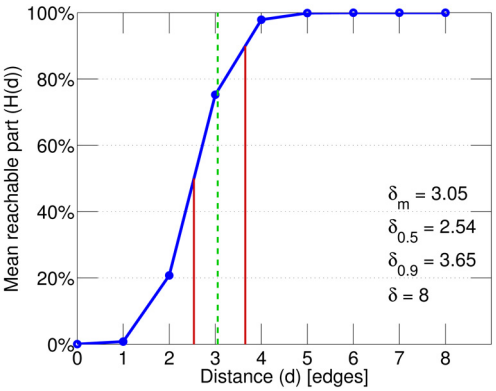
(c) *Degree distribution*



(d) *Out/in-degree comparison*



(e) *Clustering coefficient distribution*



(f) *Distance distribution*

Figure 5.4: Distributions for the dataset D_3 - University of California messages.

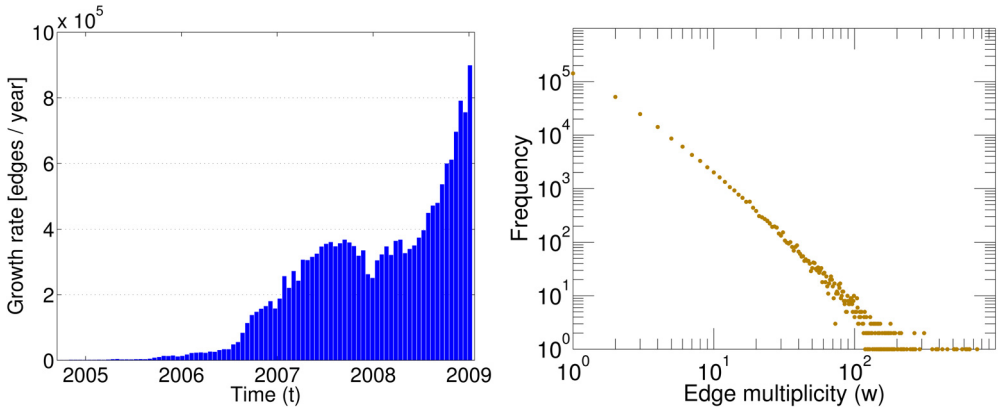
Table 5.3: Properties of the dataset D_3 - University of California messages

Property	**Value**
Period	2004-04-15 . . . 2004-10-26
Individual type	User
Event type	Message
Format	Directed
Edge weights	Multiple unweighted
Size	1,899 vertices (users)
Volume	59,835 edges (messages)
Average degree (overall)	63.017 edges / vertex
Maximum degree	1,546 edges
Reciprocity	6.363816×10^{-1}
Largest connected component	1,893 vertices
Relative largest connected component	99.7%
Largest strongly connected component	1,294 vertices
Power law exponent (estimated)	2.8310 ($d_{min} = 35$)
Gini coefficient	75.4%
Clustering coefficient	5.683030×10^{-2}
Diameter	8 edges
90-percentile effective diameter	3.65 edges
Median shortest path length	3 edges
Mean shortest path length	3.05 edges
Preferential attachment exponent	0.87269 ($\epsilon = 5.8628$)

5.2.2.4 Facebook wall posts

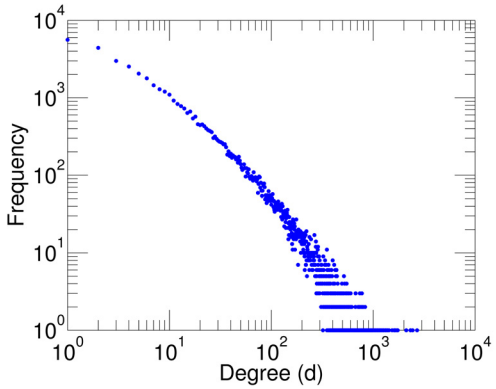
Description: This is the directed network of a small subset of posts to other user's wall on Facebook (Viswanath et al. [2009]). The nodes of the network are Facebook users, and each directed edge represents one post, linking the users writing a post to the users whose wall the post is written on. Since users may write multiple posts on a wall, the network allows multiple edges connecting a single node pair. Since users may write on their own wall, the network contains loops.

URL: http://konect.uni-koblenz.de/networks/facebook-wosn-wall

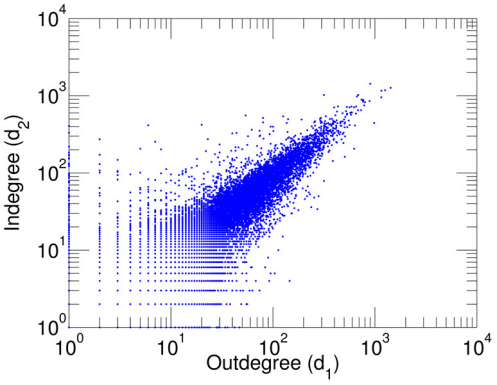


(a) *Temporal distribution*

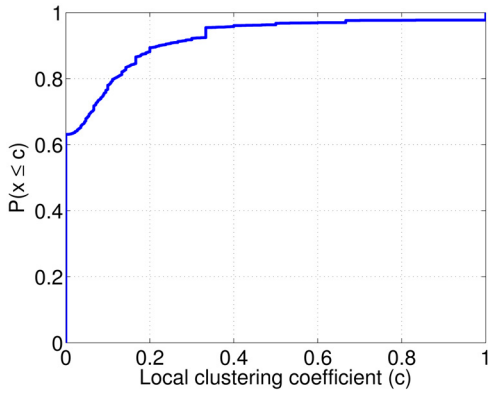
(b) *Edge multiplicity distribution*



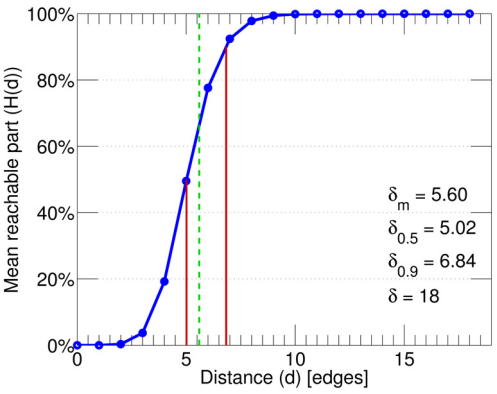
(c) *Degree distribution*



(d) *Out/in-degree comparison*



(e) *Clustering coefficient distribution*



(f) *Distance distribution*

Figure 5.5: Distributions for the dataset D_4 - Facebook user to user wall posts.

Table 5.4: Properties of the dataset D_4 - Facebook user to user wall posts

Property	**Value**
Period	2004-09-14 . . . 2009-01-22
Individual type	User
Event type	Wallpost
Format	Directed
Edge weights	Multiple unweighted
Size	46,952 vertices
Volume	876,993 edges (wallposts)
Average degree (overall)	37.357 edges / vertex
Maximum degree	2,696 edges
Reciprocity	6.248623×10^{-1}
Largest connected component	43,953 vertices
Relative largest connected component	93.6%
Largest strongly connected component	30,793 vertices
Power law exponent (estimated)	4.4310 ($d_{min} = 50$)
Gini coefficient	73.5%
Clustering coefficient	8.509687×10^{-2}
Diameter	18 edges
90-percentile effective diameter	6.84 edges
Median shortest path length	6 edges
Mean shortest path length	5.60 edges
Preferential attachment exponent	0.87210 ($\epsilon = 5.0082$)

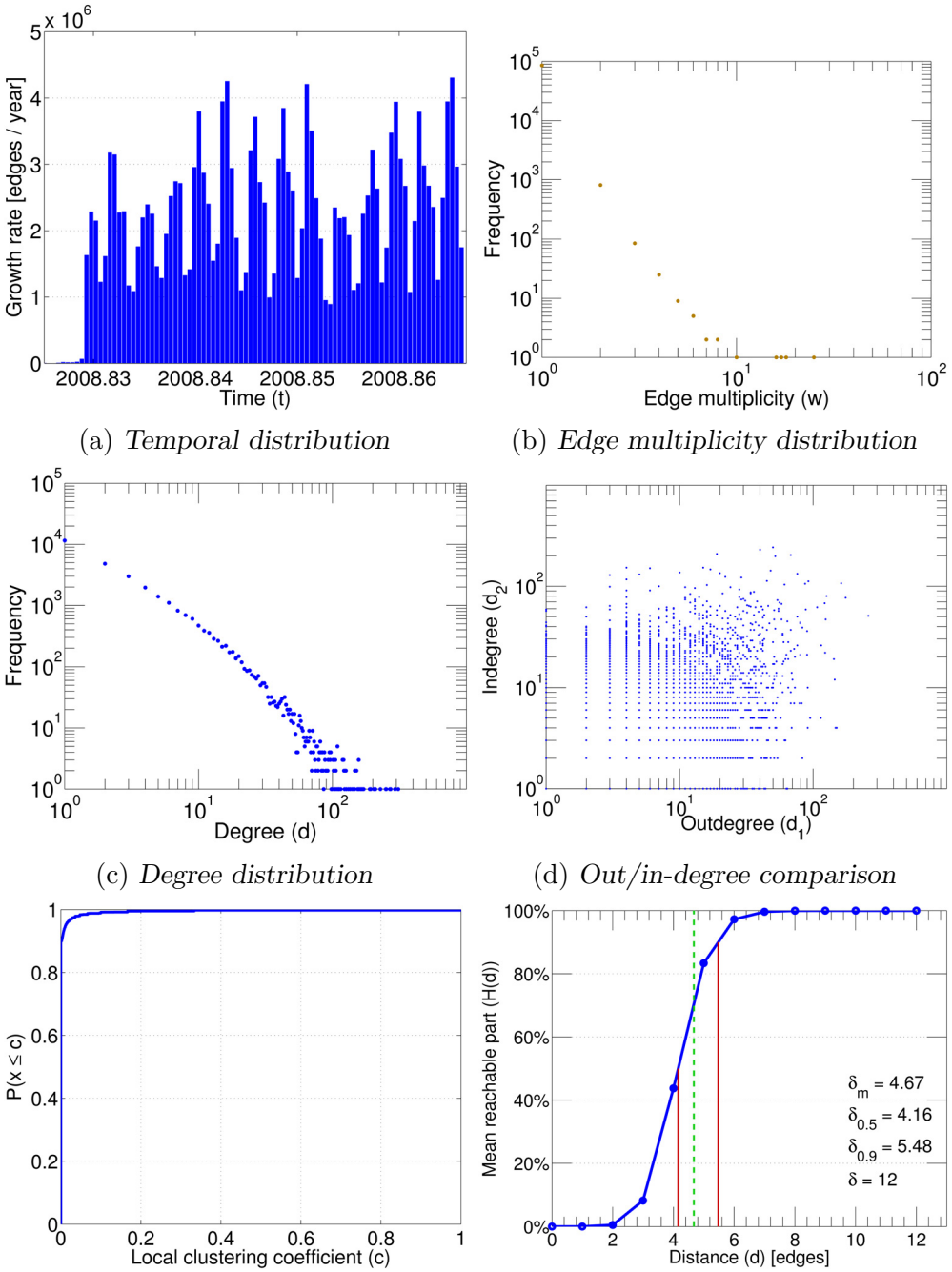
5.2.2.5 Digg reply network

Description: This is the reply network of the social news website, Digg (De Choudhury et al. [2009]). Each node in the network is a user of the website, and each directed edge denotes that a user replied to another user.

URL: http://konect.uni-koblenz.de/networks/munmun_digg_reply

5.2.2.6 Datasets - Summary

The above datasets were selected for numerous reasons briefly presented in this section. The basic reason was the different number of individuals and events in



(a) *Temporal distribution* (b) *Edge multiplicity distribution*

(c) *Degree distribution* (d) *Out/in-degree comparison*

(e) *Clustering coefficient distribution* (f) *Distance distribution*

Figure 5.6: Distributions for the dataset D_5 - Reply network of the Digg website.

Table 5.5: Properties of the dataset D_5 - Reply network of the Digg website

Property	**Value**
Period	2008-10-28 . . . 2008-11-13
Individual type	User
Event type	Reply
Format	Directed
Edge weights	Multiple unweighted
Size	30,398 vertices
Volume	87,627 edges (replies)
Average degree (overall)	5.7653 edges / vertex
Maximum degree	310 edges
Reciprocity	1.552012×10^{-2}
Largest connected component	29,652 vertices
Relative largest connected component	97.5%
Largest strongly connected component	6,746 vertices
Power law exponent (estimated)	2.6910 ($d_{min} = 16$)
Gini coefficient	63.2%
Clustering coefficient	5.598198×10^{-3}
Diameter	12 edges
90-percentile effective diameter	5.48 edges
Median shortest path length	5 edges
Mean shortest path length	4.67 edges
Preferential attachment exponent	0.50236 ($\epsilon = 3.0870$)

particular datasets. The smallest dataset in terms of number of individuals is D_1 with only 167 nodes which consists of emails exchanged between employees of a manufacturing company. Since this dataset covers the period of nine months, it has a relatively large number of events, 82,927 which makes it not the smallest one in terms of the number of events. This dataset also has the smallest shortest path median, the smallest diameter, and it is the only one that has a single connected component. Moreover, when looking at temporal properties (see Figure 5.2a), the increase in number of events in time is relatively small, so the network may be considered as a relatively stable. Unfortunately, the small number of individuals does not reveal the power law in the degree distribution, while for other networks it is observed. It is also interesting that this is the only dataset that has a strong correlation of in-degree and out-degree (see Figure 5.2d).

The second dataset is the well-known Enron email dataset - D_2. It has been used in many research studies (e.g. Kossinets et al. [2008]; Michalski et al. [2011b]; Tang et al. [2008]). It consists of the largest number of individuals (nearly ninety thousands) and the largest number of events (over one million). It reveals the power law distributions in many aspects, such as edge multiplicity (Figure 5.3b) or degree (Figure 5.3c). Due to the number of events, processing of this network takes the longest time compared to the others. In terms of temporal properties, it can also be considered as relatively stable.

The third dataset, D_3, represents the messages between users of an online community from the University of California. In terms of the number of individuals, it is the second smallest dataset, since it has 1,899 nodes. This dataset has an interesting growth rate (Figure 5.4a), since the number of edges increases to a certain period, then the growth rate drops. The diameter of the dataset is small and equals three.

The next dataset, D_4, is the user-to-user Facebook wall posts log. It is nearly as big as the Enron one, since it has about 50,000 individuals and nearly 900,000 events. One of the most interesting properties of it is that is grows in time, as Figure 5.5a shows. This property makes the task of maximizing the spread of influence more challenging, since with rapid growth the future structure of the network is more unpredictable. Similarly to the Enron case, this dataset is time-consuming.

Finally, the last dataset is taken from Digg social news website and it is a user-to-user reply sequence. It is placed in the middle of other datasets when about most properties, such as number of individuals or events. Moreover, as the dataset covers just about a two week period, in Figure 5.6a the night and day periods are clearly distinguishable. But one of the most important properties of this dataset cannot be seen from the information presented above, but in Table 5.11 - the individuals rarely interact again, so there are no stable relations. But this will be discussed later in Section 5.3.2.1.

It is worth mentioning that despite the information presented in Tables 5.1-5.5 (Edge weights - multiple unweighted), the resulting TSNs will be weighted (and directed), as presented in Definition 1. The weights mentioned in those tables refer to the weights over events, which indeed do not exist.

5.2.3 Time-limited Event Sequences

Before generating the temporal social network, a process of narrowing down the datasets was performed. The idea was to evaluate the results of seeding strategies in different configurations of the datasets. Since the nature of social networks is often bursty (Barabási [2010]), it was interesting to see how different seeding strategies would perform regardless of the moment of when they were applied. This is why for each dataset $D_1, \ldots, D_5$ five separate testing and training TESes are generated. The procedure is as follows.

For each of the datasets $D_1, \ldots, D_5$ which in fact follow the ES definition (see Section 2.2), five TESes have been generated (see Section 2.4 for details), denoted as $TES^{Tr}_{D_z,T^{Tr}_u}$. Here, D_z corresponds to the dataset identifier, $1 \leqslant z \leqslant 5$, and the value of shift index u (the subscript of T^{Tr}_u) indicates one of the scenarios presented in Table 5.6 and Figure 5.7. The period T in the table represents the whole period covered by the dataset D_z.

As the Table 5.6 indicates, for each dataset there were five TESes created, each of them consisting of events covering 60% of period covered by the dataset, but they differ by the shift of $0\%, 5\%, 10\%, 15\%, 20\%$ from the beginning, respectively (see Figure 5.7). In all cases the closure types were the same, $closure_l$ = '[' and $closure_u$ = ')'. The superscript Tr in the expression $TES^{Tr}_{D_z,T^{Tr}_u}$ means that for the influence maximization problem, these TESes will be used for *training* purposes, i.e. the events contained by TES will be the source data for choosing seeds which will be influenced in the first step (for the sake of clarity Tr is the abbreviation for training).

Moreover, for each $TES^{Tr}_{D_z,T^{Tr}_u}$ introduced above there was one additional TES generated, denoted as $TES^{Te}_{D_z,T^{Te}_u}$. This time-limited event sequences has properties presented in Table 5.7 for each u. This TES refers to the testing part of the datasets where particular seeding strategies will be evaluated and the abbreviation Te stands for testing.

In Figure 5.7 all the configurations of different values of shift u are presented, the brackets symbolise the values of $closure_l$ (the beginning of a particular TES period) and $closure_u$ (the end of TES period).

Table 5.6: Training time-limited event sequences generated from the datasets $D_1, \ldots, D_5$.

u	t_p^{Tr}	$t_{p'}^{Tr}$	$closure_l$	$closure_u$
1	$min(T) + 0.00 * (max(T) - min(T))$	$min(T) + 0.60 * (max(T) - min(T))$	'['	')'
2	$min(T) + 0.05 * (max(T) - min(T))$	$min(T) + 0.65 * (max(T) - min(T))$	'['	')'
3	$min(T) + 0.10 * (max(T) - min(T))$	$min(T) + 0.70 * (max(T) - min(T))$	'['	')'
4	$min(T) + 0.15 * (max(T) - min(T))$	$min(T) + 0.75 * (max(T) - min(T))$	'['	')'
5	$min(T) + 0.20 * (max(T) - min(T))$	$min(T) + 0.80 * (max(T) - min(T))$	'['	')'

Table 5.7: Testing time-limited event sequences generated from the datasets $D_1, \ldots, D_5$.

u	t_p^{Te}	$t_{p'}^{Te}$	$closure_l$	$closure_u$
1	$min(T) + 0.60 * (max(T) - min(T))$	$min(T) + 0.80 * (max(T) - min(T))$	'['	']'
2	$min(T) + 0.65 * (max(T) - min(T))$	$min(T) + 0.85 * (max(T) - min(T))$	'['	']'
3	$min(T) + 0.70 * (max(T) - min(T))$	$min(T) + 0.90 * (max(T) - min(T))$	'['	']'
4	$min(T) + 0.75 * (max(T) - min(T))$	$min(T) + 0.95 * (max(T) - min(T))$	'['	']'
5	$min(T) + 0.80 * (max(T) - min(T))$	$min(T) + 1.00 * (max(T) - min(T))$	'['	']'

5.2.3.1 Properties of TESes

Table 5.8 shows what particular TESes look like when about the number of individuals and events in training and testing periods. The goal of this table is to show the diversity across the datasets. The value of an average number of events per individual in training TES was normalized to the length of the testing period for comparison purposes. Moreover, for the testing set information about the number of new nodes that joined the network after the training period is provided.

This table provides a number of facts about the TESes and allows comparison between them in terms of size, stability and activity in event sequences. This information extends the information about the datasets provided in Section 5.2.2. For instance, the test that is built using the dataset D_1 is relatively stable in terms of individuals, since for all the periods' configurations no new node joins in the testing period. It contrasts with TESes using datasets D_2 and D_4 where in the testing periods new nodes are observed. Moreover, it is noted that the highest activity of individuals happens in TES using D_1, since they are involved in 10-60 times more events than individuals in other datasets, mainly because of the length covered by the dataset.

More detailed discussion on TESes generated here is presented in Section 5.3.4.1.

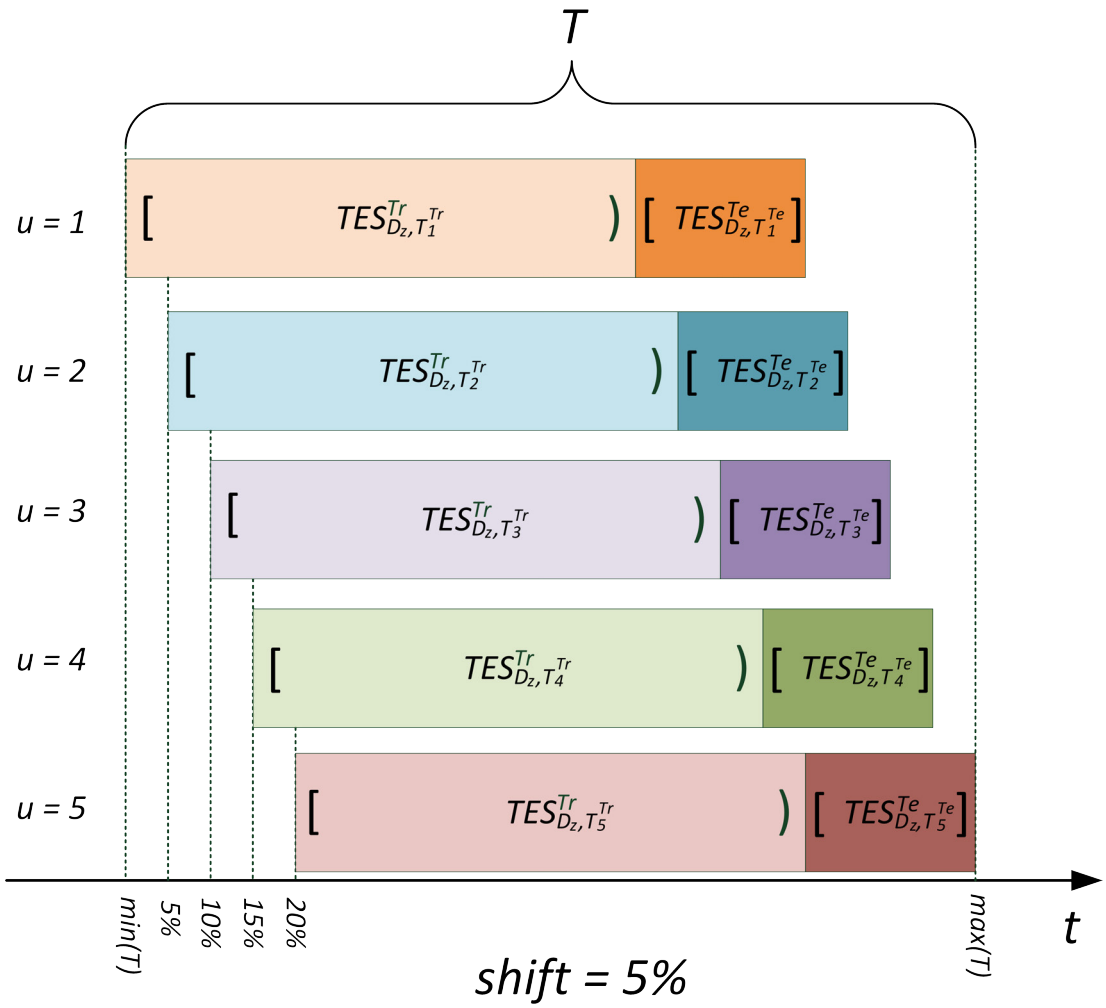


Figure 5.7: Five possible configurations of $TES^{Tr}_{D_z,T^{Tr}_u}$ and $TES^{Te}_{D_z,T^{Te}_u}$, depending on the value of u.

Table 5.8: Number of individuals and events in training and testing TESes.

Dataset index z	Shift index u	$TES^{Tr}_{D_z,T^{Tr}_u}$			$TES^{Te}_{D_z,T^{Te}_u}$			
		No. of individuals	No. of events	Avg. no. of events per individual	No. of individuals	No. of new individuals	No. of events	Avg. no. of events per individual
1	1	166	50,391	101.19	138	0	16,147	117.01
1	2	165	49,414	99.83	138	0	15,434	111.84
1	3	157	48,694	103.38	136	0	14,652	107.74
1	4	156	47,377	101.23	138	0	15,615	113.15
1	5	156	46,924	100.26	140	0	16,389	117.06
2	1	30,331	346,339	3.81	46,434	34,450	506,146	10.9
2	2	36,130	459,816	4.24	55,697	42,391	579,276	10.4
2	3	47,165	630,836	4.46	51,997	38,229	495,691	9.53
2	4	54,272	734,687	4.51	45,298	32,377	406,560	8.98
2	5	64,629	849,244	4.38	35,683	22,325	294,641	8.26
3	1	1,786	54,870	10.24	477	55	3,170	6.65
3	2	1,759	54,748	10.37	470	72	3,135	6.67
3	3	1,707	47,296	9.24	395	53	2,653	6.72
3	4	1,542	35,807	7.74	368	65	2,325	6.32
3	5	1,389	24,522	5.88	368	63	1,795	4.88
4	1	12,045	149,227	4.13	21,303	10,452	286,537	13.45
4	2	14,323	220,144	5.12	24,037	11,446	290,280	12.08
4	3	17,150	297,869	5.79	28,246	13,492	296,634	10.5
4	4	19,904	370,524	6.21	34,589	17,692	337,602	9.76
4	5	22,476	434,814	6.45	43,655	24,466	441,229	10.11
5	1	21,132	47,768	0.75	11,331	4,444	17,857	1.58
5	2	22,638	53,510	0.79	11,162	4,107	17,595	1.58
5	3	23,437	56,714	0.81	12,112	4,491	19,488	1.61
5	4	23,594	56,776	0.8	12,715	4,809	20,473	1.61
5	5	23,608	56,469	0.8	13,456	5,153	22,002	1.64

5.2.4 Temporal Social Networks Configuration

In Section 5.2.3 the TESes were defined and they will be the base for generating Temporal Social Networks TSNs (see Section 2.5 for definition). Initially, for each of the TESes ten to twelve TSNs were created consisting of a different number SNs. These are denoted in the following way: $TSN^{K}_{D_z,T_u}$, where D_z refers to the dataset and z is the dataset index, T_u is the period and u indicates the shift (see Section 5.2.3). Finally, K refers to the number of time windows the TSN consists of (please note the different placement of index K to form a superscript for aesthetics purposes). Moreover, all the TSNs were built from non-overlapping consecutive time-windows of the same size, i.e. a TES was split in K time windows of equal size where none of them contain events of other TESes and the whole period T_u is covered by SNs being part of this TSN (see Figure 2.3a). This approach comes from the idea of the temporal model for the spread of influence described in in 6.3.2.

Created TSNs are directed and weighted. The weight over edges is calculated as presented in Definition 2.3.1, i.e. $w_{ij} = \frac{n^e_{ij}}{n^e_i}$ is the importance (weight, strength) of the relationship between individuals, such that n^e_{ij} is the number of events ev_{ijkl} from v^e_i to v^e_j in ES (regardless k, l) and n^e_i is the number of all events initiated by v^e_i (outgoing from).

For instance, $TSN^{8}_{D_1,T^{Tr}_2}$ is a temporal social network generated from the manufacturing company emails event sequence based on dataset D_1, period T^{Tr}_2 is the training period which was shifted by 5% and the index in superscript 8 indicates that this temporal social network consists of 8 social networks.

It has been decided to evaluate how different network properties would look when the TSN consisted of a different number of SNs. But, instead of creating networks consisting of $1, 2, 3, \ldots, K$ time windows, another approach was taken, where K was increased as a power of 2. So, K varied from 1 (time-aggregated) network to 2048, i.e. 2^{11}.

Table 5.9 shows the length of the period for each TES used for generating TSNs as well as which TSNs were generated from it. In this scenario 270 TSNs were created - 54 configurations of TSNs times five possible shift values. In the case of datasets D_1, D_3 and D_5, TSNs consisting of 1024 and 2048 were not

generated, since they covered too short periods in each component SNs.

Table 5.9: Information about which TSNs were generated for each TES.

Dataset index z	Length (days)	K											
		1	2	4	8	16	32	64	128	256	512	1024	2048
1	162.6	•	•	•	•	•	•	•	•	•	•	-	-
2	808.8	•	•	•	•	•	•	•	•	•	•	•	•
3	116.4	•	•	•	•	•	•	•	•	•	•	-	-
4	954.6	•	•	•	•	•	•	•	•	•	•	•	•
5	9.6	•	•	•	•	•	•	•	•	•	•	-	-

Moreover, similarly to the case with TESes (see Section 5.2.3), testing TSNs were generated. They are denoted similarly, i.e. $TSN^{K'}_{D_z,T^{Te}_u}$, but the period T_u will now refer to the testing TESes, so it is superscripted Te. Those TSNs are of the same type and were built from non-overlapping consecutive time-windows of the same size, but in contrast to the training TSNs, K' always equals 10, i.e. the testing scenario assumes that the temporal social network consists of ten time windows (refer to Section 6.3.2 for the details).

5.2.5 Experimental Environment

5.2.5.1 Framework for Temporal Social Networks and Social Influence

To conduct all the experiments author of this dissertation built a framework for processing datasets, creating TESes and TSNs, and performing simulations of spread of influence. In this section the functionalities of the environment are presented, while the next section describes all the software and hardware that are used.

Importing datasets

The framework is able to process all the timestamped datasets that are available at KONECT - The Koblenz Network Collection[1]. After pointing to the appropriate datafile, it downloads the dataset, extracts it and imports it into the database. As for now the only information stored is the sender,

[1]http://konect.uni-koblenz.de/

recipient, time-stamp of an event and (optionally) the sender and recipient description. The data is stored in the form of Event Sequences and it is possible to obtain a Time-limited Event Sequence from it when providing the period being studied (see Sections 2.2 and 2.4 for details). For both, ES and TES, the framework can provide a number of statistics regarding individuals and events, such as the number of individuals, number of events, distribution of events in time etc.

Temporal Social Networks

By basing either on ES or TES this framework is capable of generating Temporal Social Networks (see Section 2.5) in any of the scenarios presented in Figure 2.3. Moreover, it is possible to explore new types of Temporal Social Networks where, instead of providing the time-windows period, researcher can focus on events, e.g. create TSN with windows of an equal or unequal number of events. This area has not been explored scientifically yet, but the author of this dissertation believes that it has great potential. Generated TSNs can be exported to flat file or to database, depending on the future use of them. Similarly to the case of ES/TES, the framework provides a number of statistics either for the whole TSN or for particular SNs it consists of. One can find the following: the number of nodes, number of events, structural measures such as: degree, betweenness, closeness, clustering coefficient, the number of weakly or strongly connected components etc.

Social Influence

In the framework one model of social influence is currently implemented, i.e. Linear Threshold model (see Section 3.4.3.2). It assumes that the weights over edges are the influence weights b and the threshold θ can be provided either as fixed for all nodes, assigned individually for each node or drawn from a Gaussian distribution. Apart from that it is required to provide a seed set for the influence process and this is possible in one of three ways: either by providing the exact set of nodes or by providing the budget c as a percentage of nodes found in the TSN (rounded up) and the algorithm of choosing seeds. Algorithms are of the following types:

- *random* - choosing nodes at random
- $random_{freq}$ - selecting nodes based on their frequency of occurrence in particular time windows
- *greedy* - choosing nodes with the greedy algorithm (see Section 4.4.3) for the social network identifier provided
- *rank* - choosing nodes from ranks generated by measures described in Section 6.6.5, either taking c percent of nodes from the top or bottom of the rank

The last set of information needed to start the influence process is the time window in which the process should start and how many time windows it should last for (stop condition).

Even though the framework was designed to work with temporal social networks, it is also capable of generating time-aggregated ones and running the influence process there, since it is only a matter of generating an appropriate network with a single time window instead of multiple windows.

5.2.5.2 Parallel Computing

For two operations within the framework it was essential to implement parallelism to increase the speed of computations. These are the following:

- *greedy algorithm* - finding nodes that maximize the influence in a greedy search is not a trivial task, luckily it can be run in parallel. For each evaluated node separate process were run to check this node in terms of influencing others. Then results for all the nodes were compared to choose the best node and add it to the seed set.
- *spread of influence* - after having an initial seed set selected, the process of influence starts. In the LT model each node verifies what weighted fraction of its neighbours is influenced, and based on that and the influence threshold θ, whether it becomes influenced or not. In the framework this task was also run in parallel for each node in every time window.

Implementing the parallelism allowed the running of the experiments in a supercomputer environment (Supernova cluster), which is established in Wrocław Centre for Networking and Supercomputing WCSS[1]. The calculations were run under WCSS grant number 177.

5.2.5.3 Hardware and Software

This framework was implemented in R programming language[2] (Team et al. [2005]) - version 3.1.0. The package for graph processing is igraph[3] (Csardi and Nepusz [2006]). For the parallelism the *foreach* and *doMC* packages were used and for the database connection - RMySQL.

The database for storing either ESs or TSNs was established in MySQL server environment[4] - version 5.5.37. As it has already been mentioned in Section 5.2.5.2, some operations were run in parallel in a supercomputing environment. In order not to overload the database with remote requests to obtain particular SNs, the database was used only to store datasets, event sequences and temporal social networks, as well as for providing them for calculating measures by local R instance. The remote (cluster) nodes extracted the TSNs from files exported from the database and saved the results to files as well. This solution increased speed by getting rid of database connectivity from cluster environment to local database server and stability, since the cluster nodes were not dependent on external sources. The whole environment is presented in Figure 5.8.

The Supernova cluster nodes have the following configuration. Each node has two Intel Xeon X5650 2.67 GHz 6-core processors with 12MB cache and 24GB of RAM (about 22 GB effectively). The operating system is ScientificLinux[5]. For the greedy algorithm and the spread of influence the task was decomposed among multiple nodes with a given set of temporal social network nodes to be evaluated. Theoretically, a single host could evaluate 12 graph nodes concurrently. Unfortunately, as the experiments were run on temporal social networks, not time-aggregated ones, there were also memory constraints, since the networks were

[1] http://www.wcss.pl/en/
[2] http://www.R-project.org
[3] http://www.igraph.org
[4] http://www.mysql.com
[5] https://www.scientificlinux.org/

bigger. For instance, D_1 requires about 500 MB for a temporal spread of influence scenario being evaluated and D_2 - 4 GB. Here, a single node would be able to work with 12 TSNs built using D_1, but only with 5-6 TSNs built by using D_2. This is why the author of this dissertation tried to optimize the calculation time by maximizing the nodes' usage, i.e. trying to use all the available cores within a single host by placing computations for different temporal social networks within them. For instance, a single host computed 4 TSNs for D_2 (16 GB RAM, 4 cores) and 8 TSNs for D_1 (4 GB RAM, 8 cores) which resulted in using all the cores and most of the memory.

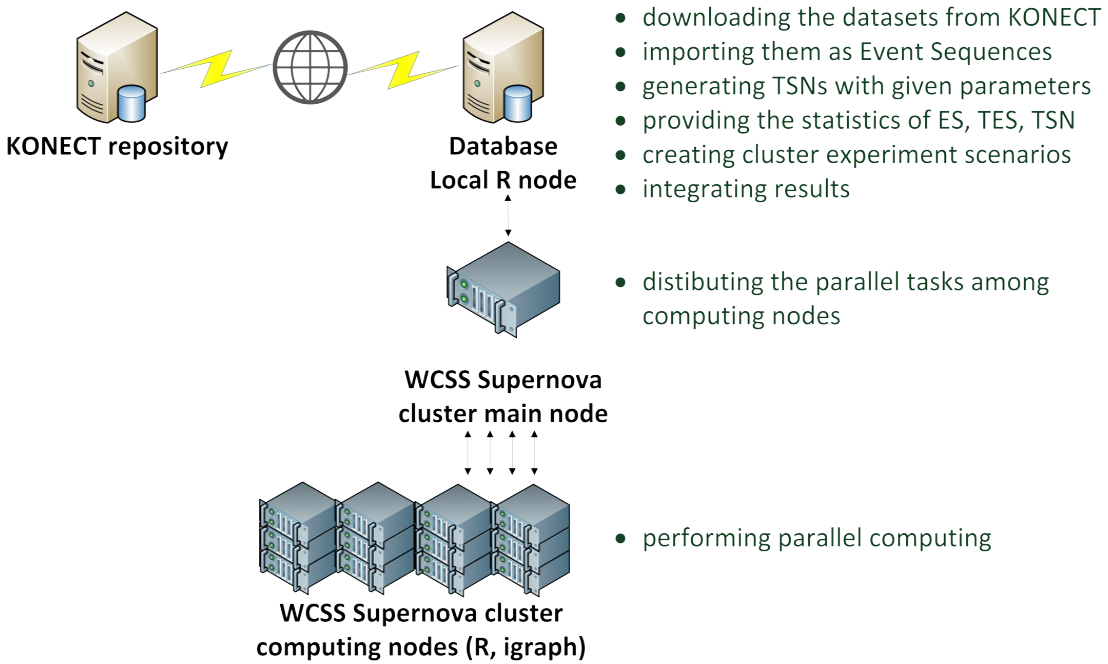


Figure 5.8: The computing environment for the developed framework.

5.3 Network Properties of Temporal Social Networks

5.3.1 Introduction

As was presented in Section 5.2.4, a number of Temporal Social Networks were generated. The purpose of this part of the dissertation is to analyse how the way they were generated influences the network structural measures and general properties, such as the number of nodes and edges and their overlap in particular time windows. The assumption is that by splitting the datasets so strongly (from 1 to 2048 time windows) some structural features may appear. Moreover, it is assumed that each network has its pattern which is strongly connected with the network type. For instance, a manufacturing company has its working hours and daily habits of employees regarding who contacts whom. If we begin to split the network, maybe at some point these regularities will be broken due to the fact that they will be simply "cut" in half.

This section is organized as follows. In the next subsection general network properties of obtained TSNs are presented, while in Section 5.3.2 the author of this dissertation focuses on structural network measures. Both parts are required to look at the temporal networks in terms of finding opportunities to exploit particular sizes of time windows to maximize the spread of influence, which will be presented in Chapter 6. Moreover, they present some phenomena which may go against intuition when considering increasing the number of network cuts.

5.3.2 Stability of Network Properties

5.3.2.1 Similarities Between Time Windows

The objective of this section is to show how social network properties change if splits become bigger and bigger. Moreover, the idea is to obtain a deeper understanding of analysed temporal social networks in terms of their characteristics, since each dataset has different properties (see Section 5.2.2). This characteristic will be now studied at a different level.

While Table 5.9 presents of each networks how many temporal social networks

are generated of each kind, Table 5.10 shows the length of a single time window in particular configurations. Naturally, this implies that some properties of TSNs will be different for differing granularity, but, what is even more important, is that some interesting facts regarding the behaviour of individuals (nodes) are also revealed.

Table 5.10: The length of a single time window for all generated TSNs

Dataset index z	Length (days)											
	K											
	1	2	4	8	16	32	64	128	256	512	1024	2048
1	162.6	81.3	40.65	20.33	10.16	5.08	2.54	1.27	0.64	0.32	-	-
2	808.8	404.4	202.2	101.1	50.55	25.28	12.64	6.32	3.16	1.58	0.79	0.39
3	116.4	58.2	29.1	14.55	7.28	3.64	1.82	0.91	0.45	0.23	-	-
4	954.6	477.3	238.65	119.33	59.66	29.83	14.92	7.46	3.73	1.86	0.93	0.47
5	9.6	4.8	2.4	1.2	0.6	0.3	0.15	0.08	0.04	0.02	-	-

Firstly, it is interesting to see how stable the individuals in the TSNs are, i.e. whether the rotation of them between particular time windows is high or not. This is an important issue, since in contrast to time-aggregated networks where network persistence of a node is fixed, the temporal social networks let the nodes move themselves and change their neighbourhood. This property may be beneficial to the process of social influence, because if a node is influential and regularly changes its neighbourhood, it is possible that it will influence more nodes. Naturally, it is dependent on influence weights b and threshold level θ, but this kind of node behaviour may support the general process outcomes. For static networks if the process is stalled, i.e. no more nodes can be influenced, nothing more can be done and the process ends. In the temporal case with interchanging links there is still the chance that some new configuration will end up with newly influenced nodes.

Table 5.11 presents how similar in terms of nodes and edges are particular consecutive SNs of selected TSNs. In this case the analysed TSNs are those consisting of eight time windows and without any shift - $TSN^{8}_{D_z,T_1^{Tr}}$ - for each z. This table shows the ratio of nodes (top part) and edges (bottom part) from a previous time window that exists in the next time window. For instance, for the first row and first cell, 0.98 means that 98% of nodes (or edges) from the first time window exist in the second time window. Moreover, the last time window is compared with the first one showing how much the network changed in time since the beginning (last column).

As one can see in Table 5.11, there are three different categories of networks here. Firstly, the most stable one, where almost all nodes that exist in all time windows and edges between are relatively stable. Only one TSN suits here, namely manufacturing company e-mails ($z = 1$). The stability of nodes is over 90% and about $50 - 60\%$ edges persist in the next time window. The second group is characterized by the decreasing percentage of similar nodes when moving towards the newer (later) time windows. This reflects TSNs generated from D_3 and D_5. Lastly, the number of similar nodes increases in datasets D_2 and D_4. However, when looking at the last column it is seen that for the dataset D_3 nodes from the newest time window are different to those at the beginning (0.17).

Interesting conclusions may be drawn from the comparison of edges' similarity.

Table 5.11: The similarity of social networks in $TSN^{8}_{D_z,T_1^{Tr}}$ - for nodes and edges.

Dataset index z	Similarity							
	2 & 1	3 & 2	4 & 3	5 & 4	6 & 5	7 & 6	8 & 7	8 & 1
	Nodes							
1	0.98	0.94	0.98	0.93	0.97	0.9	0.93	0.84
2	0.39	0.65	0.44	0.54	0.53	0.55	0.56	0.76
3	0.77	0.72	0.64	0.39	0.53	0.54	0.5	0.17
4	0.47	0.47	0.68	0.71	0.75	0.86	0.85	0.86
5	0.81	0.42	0.3	0.37	0.38	0.32	0.35	0.42
	Edges							
1	0.55	0.56	0.53	0.49	0.46	0.63	0.63	0.44
2	0.17	0.48	0.19	0.19	0.26	0.29	0.29	0.41
3	0.2	0.16	0.08	0.07	0.09	0.15	0.19	0.01
4	0.29	0.29	0.33	0.31	0.27	0.38	0.39	0.32
5	0	0	0	0	0	0	0	0

For three TSNs, built from datasets D_1, D_2 and D_4 the edges are similar on the average of 30-40%. However for two TSNs - D_3 (University of California) and 5 (Digg) the similarity of edges is really small and even 0% for the Digg dataset. This means that in this TSN users do not form stable relationships and change them often for this kind of network split.

The above comparison is presented for $K = 8$, but in Figures 5.9 and 5.10 there is presented the averaged similarity for all values of K and all TSNs in the case of nodes and edges, respectively. Again, here the shift was 0%. For most cases the similarity decreases if K increases, but there are also exceptions from this rule, for nodes and edges case. Interestingly, if the similarity for nodes increases, the same happens to edges (see $z = 2$ and $z = 3$).

5.3.2.2 Core Nodes

One of strategies of seeding in temporal social networks is to choose nodes that simply exist in the network and, ideally, often change their neighbourhood. In the previous section it was presented that most of the analysed TSNs is characterised by fluctuation of nodes, i.e. the nodes and edges are not persistent from time window to time window. However, it would be interesting to see whether there

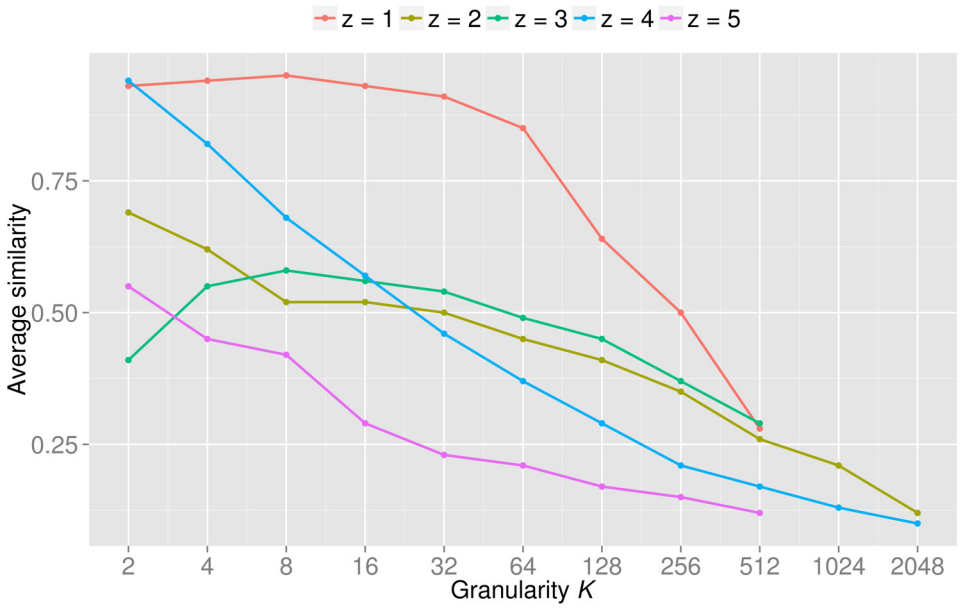


Figure 5.9: Average similarity (nodes) between time windows for a given K.

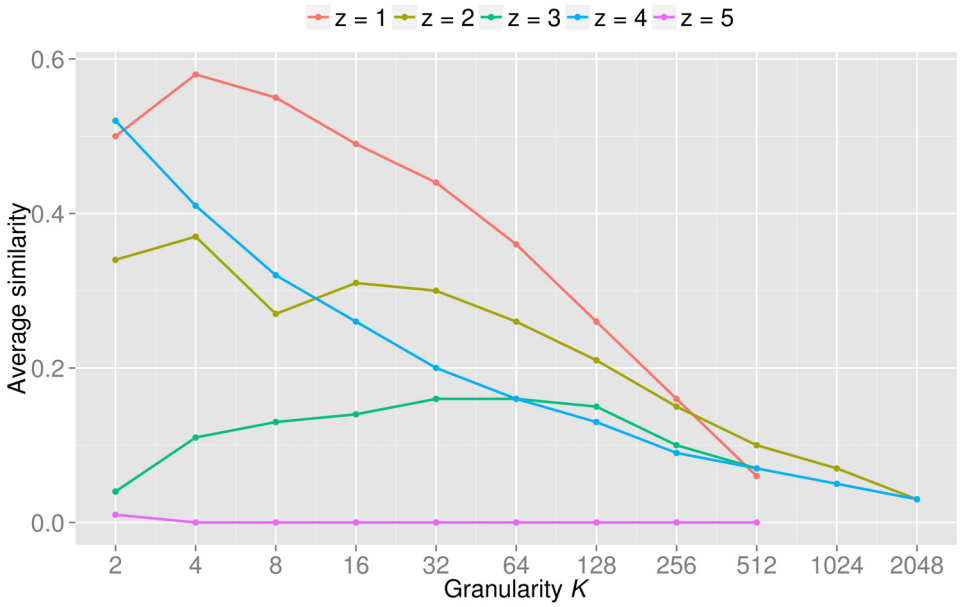


Figure 5.10: Average similarity (edges) between time windows for a given K.

are any nodes that are the *core* of the temporal social network, i.e. they exist in every time window and may potentially be used as seeds in the spread of influence process. The best option would be if these nodes exist in each time window and swap the neighbourhood. This seeding strategy will not be evaluated further in the thesis, but the stability of nodes in the temporal social network is one of the important factors in the analysis and may potentially be used in future work. The relevance to this work is that thanks to this analysis it will be seen whether is it even possible to count on stable nodes in the network or rather it should be assumed that the best option to maximize the spread of influence is to count on new nodes becoming influenced, even if they will not influence others, because they will simply not appear in the future. In contrast to the results presented in the previous subsection, here the results were averaged across all the shifts u, since this process was very prone to the starting point, i.e. if the starting window is in the middle of the night, the number of initial nodes is small and by averaging the results it was possible to avoid this problem. The number of *core* nodes for each TSN is presented in Figure 5.11.

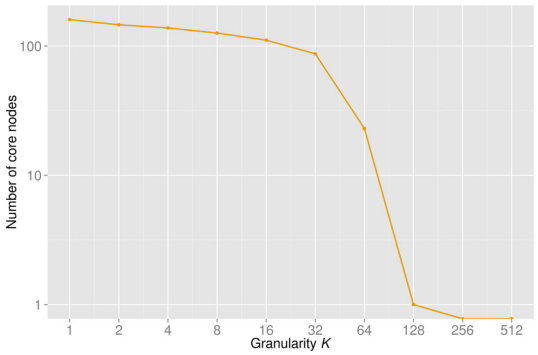
As it is presented in Figure 5.11, most of the networks here follow the power law (Barabási and Albert [1999]) and for each network there is a subset of individuals that may be interesting in terms of being hubs for information diffusion or the spread of influence. They guarantee that the budget c spent on influencing them at least will not vanish instantly. Interestingly, when comparing this figure with Figure 5.10, even for Digg TSN ($z = 5$), these nodes do not build stable relationships, since there is no overlap between edges.

5.3.3 Stability of Structural Measures

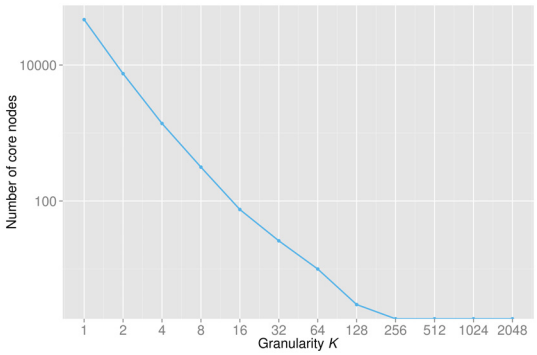
The next part of this chapter focuses on comparing generated temporal social networks in terms of structural measures to see how the granularity influences those measures and whether some crucial points exist which may be further exploited.

5.3.3.1 Methodology

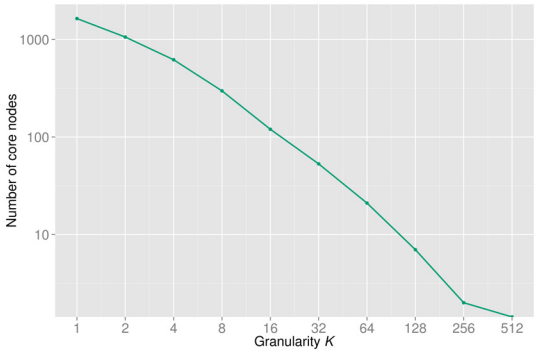
Since for each ES there were five TESes created and for each of them a number of TSNs were generated (see Section 5.2.3 and Section 5.2.4 for details), a



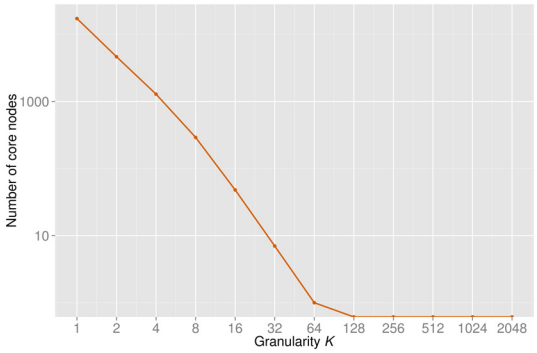
(a) *Manufacturing company*, $z = 1$



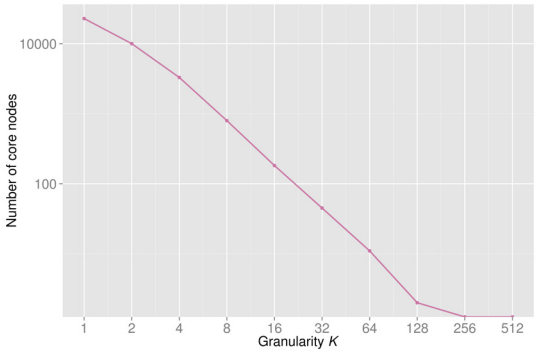
(b) *Enron*, $z = 2$



(c) *University of California*, $z = 3$



(d) *Facebook*, $z = 4$



(e) *Digg*, $z = 5$

Figure 5.11: Core nodes of the TSNs - nodes that appear in every time window

methodology has to be selected to compare these TSNs. Firstly, it is studied how similar particular TSNs are for different shifts u in terms of structural measures and whether it is possible to introduce any simplifications in further comparison process. Then, a number of network measures were introduced and TSNs consisting of a different number of time windows were compared against each other. However, firstly the measures are introduced.

5.3.3.2 Measures

Typical SNA techniques start with evaluating social networks in terms of a number of structural measures which are helpful in understanding how the network looks and which type it is. This understanding allows either the alignment of the network to one of the network models, such as Barabási-Albert (Barabási and Albert [1999]), Watts-Strogatz (Watts and Strogatz [1998]), Erdos-Rényi (Erdos and Rényi [1960]) or simply to have some idea of its characteristics and shape.

As it was presented in Section 2.6, the representation of TSNs used in this dissertation is one of many possible. Another way of representing temporal networks are contact sequences or interval graphs (Holme and Saramäki [2012]), but the representation used in this thesis has one of the advantages over the two approaches mentioned. Namely, since the DNA is in its early stage, there are not many temporal network measures that are considered to be established among researchers. This is one of the reasons why this ES $\rightarrow$ TES $\rightarrow$ TSN approach was selected - to make the networks comparable by using scientifically-agreed methods of SNA. The TSN, consisting of multiple SNs, provides an opportunity to compare individual SNs by using those measures and, at a global level, it will allow to give a broader view on the whole TSN.

After justifying the method of creating TSNs, the following measures were used for evaluation:

Betweenness centrality

Betweenness centrality shows to what extent a node is between other nodes. Members with a high value of betweenness are very important to the network, since others can connect with each other only through them. It can be calculated only for undirected relationships by dividing the number of short-

est geodesic distances (paths) from node v_i to v_j by the number of shortest geodesic distances from v_i to v_j that pass through member v_k (Carrington et al. [2005]; Musiał et al. [2009]). It is calculated as follows:

$$BC(v_x) = \frac{\sum_{v_i, v_j \in V, i \neq j} p_{v_i,v_j}(v_k)}{p_{v_i,v_j}}, \tag{5.2}$$

where $p_{v_i,v_j}(v_k)$ denotes the number of shortest paths between v_i and v_j going through v_k and p_{v_i,v_j} indicates the number of shortest paths between v_i and v_j.

In-degree centrality

In-degree centrality of a node v_i indicates how many first level neighbours point towards v_i, i.e. how many edges are directed from the neighbours of v_i to this node. This measure is also called degree prestige, since it reflects the measure that is not built by the node itself, but by others. In other words, the most prominent people are those who received more nominations from members of the community (Alexander Jr [1963]).

Out-degree centrality

In contrary to in-degree centrality, out-degree centrality focuses on the outgoing edges from a node and indicates how well a node is communicating with others. Users who communicate with a greater number of people obtain a greater out-degree centrality value. Those nodes are recognized by other network members as a crucial cog that occupies a central location in a network (Wasserman and Faust [1994]).

Total degree centrality

Total degree centrality is the sum of in-degree centrality and out-degree centrality. Since in-degree and out-degree measures reflect directed networks only, total degree centrality may also be applied to undirected networks.

Closeness centrality

The closeness centrality, in contrast denotes how close a node is to all the others within the social network. Its main idea is that the member takes

the central position if they can quickly contact other individuals in the network. This measure emphasizes quality (position in a network) rather than quantity (number of links, like in a degree measure). The member with high closeness centrality is a good propagator of ideas and information (Bavelas [1950]; Musiał et al. [2009]).

$$CC(v_x) = \sum_{i \neq x} \frac{1}{d(v_i, v_x)}, \tag{5.3}$$

where $d(v_i, v_x)$ denotes the function describing the distance between nodes v_i and v_x.

Clustering coefficient

The local clustering coefficient of a vertex in a graph quantifies how close its neighbours are to being a clique (complete graph, see Watts and Strogatz [1998]). So it may be defined as a completeness of the nodes neighbourhood in terms of edges. For directed graphs it is given a form:

$$C(v_x) = \frac{|\{e_{jk} : v_j, v_k \in N, e_{jk} \in E\}|}{k_{v_x}(k_{v_x} - 1)}, \tag{5.4}$$

where k_{v_x} represents the number of neighbours of a node v_x, so $k_{v_x}(k_{v_x} - 1)$ denotes the number of possible links that may exist between v_x neighbours.

Density

Density is the generalization of a clustering coefficient. It measures the ratio of actual edges in a social network to all the possible edges in the network (Wasserman and Faust [1994]):

$$D(SN) = \frac{2|E|}{|V|(|V| - 1)}. \tag{5.5}$$

Connected components

The last information indicates how connected the graph itself is. A connected component of an undirected graph is a subgraph in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in the supergraph. A graph that is itself connected

has exactly one connected component, consisting of the whole graph. In this thesis, strongly connected components are analysed (see Hopcroft and Tarjan [1971]).

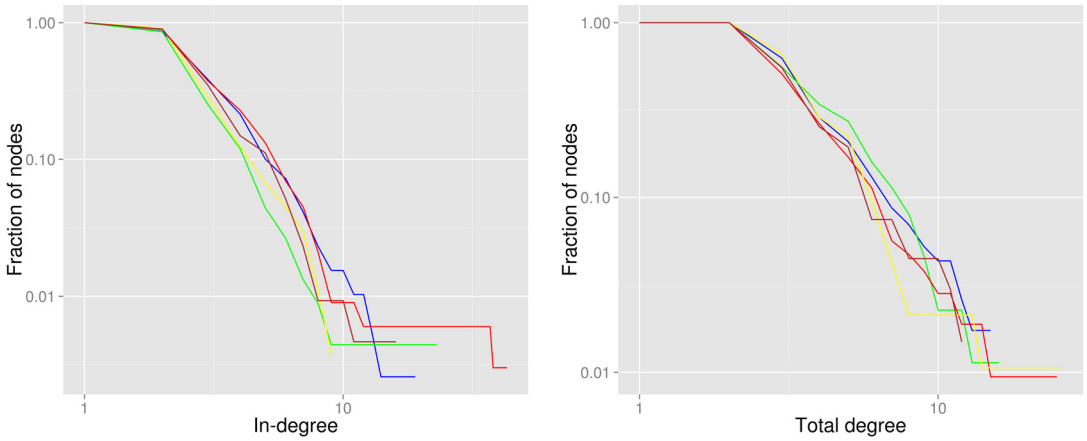
5.3.3.3 Results

The network measures are compared at two levels - firstly it is compared whether the shift introduces slight or huge differences in the distributions of measures in particular time windows. This analysis is performed for selected time windows for all networks always taking the same time window, i.e. if the first time window was taken as $TSN^8_{D_1,T_1^{Tr}}$, the comparison is being made for the first time window of $TSN^8_{D_1,T_2^{Tr}}$, $TSN^8_{D_1,T_3^{Tr}}$, $TSN^8_{D_1,T_4^{Tr}}$ and $TSN^8_{D_1,T_5^{Tr}}$ etc. Since theoretically for all the measures above and all temporal social networks there are $\sim 8 * 5 * \sum_{i=0}^{11} 2^i$ comparisons, only selected windows are compared. Results reveal that actually the shift in the datasets does not introduce major changes in distributions, as presented in Figure 5.12.

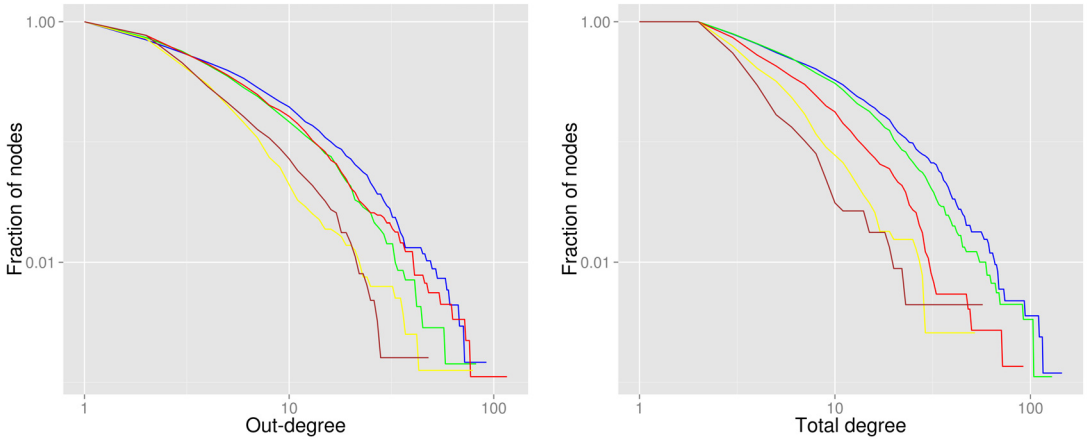
Next, it is evaluated how particular network measures change in relation to K, i.e. to the number of windows in TSN. The only way to achieve it is to present how the average values of these measures change when K is changing. To do so, for each TSN the average value of a particular measure is calculated and then it is averaged across all shifts, and presented also with the standard deviation. The results are presented in Figures 5.13-5.17, for each dataset $D_1, \ldots, D_5$. Those figures present betweenness, total degree, closeness, number of connected components and clustering coefficient altogether with density.

5.3.4 Discussion

The above presented analysis of temporal social networks properties and measures of SNs which the TSNs consists of reveals a number of interesting facts. This section summarizes and discusses them in order to foster a deeper understanding of the TSNs as well as highlighting the points of interest for in-depth analysis.

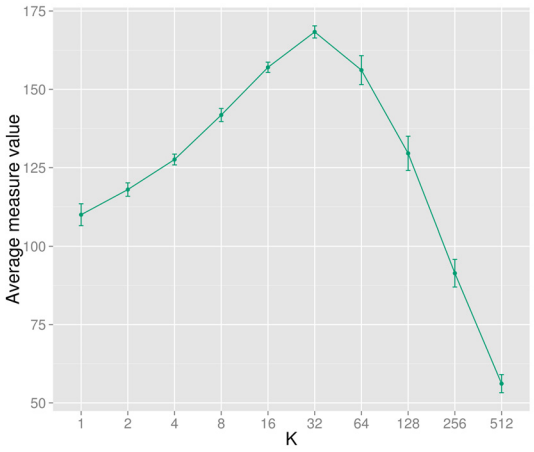


(a) *Enron:* $z = 2$, $K = 64$, *window* $= 9$ (b) *UoC :* $z = 3$, $K = 16$, *window* $= 22$

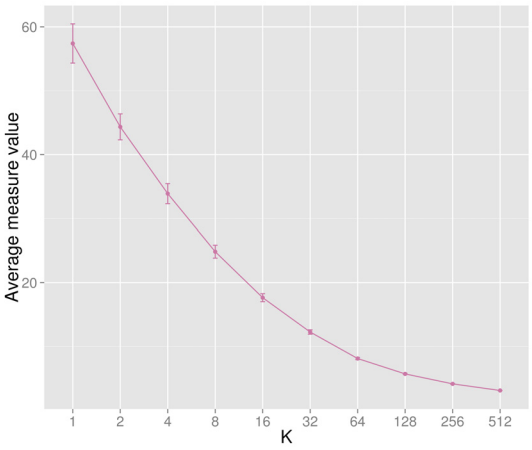


(c) *Facebook:* $z = 4$, $K = 4$, *window* $= 3$ (d) *Digg:* $z = 5$, $K = 8$, *window* $= 5$

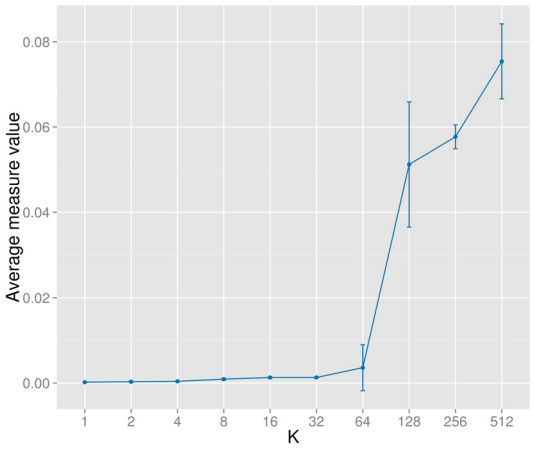
Figure 5.12: Degree cumulative distributions for selected windows of chosen TSNs and all combinations of shifts u (blue - 1, green - 2, red - 3, yellow - 4, brown - 5).



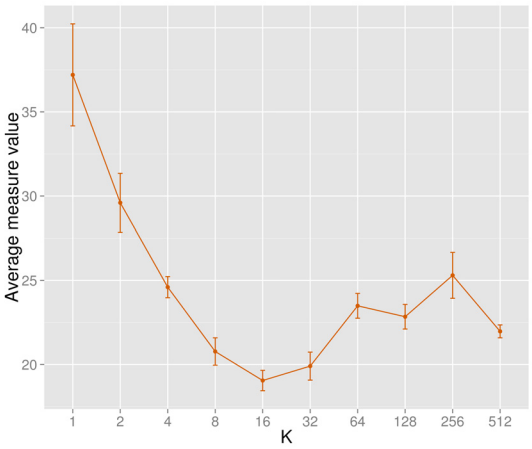
(a) *Betweenness*



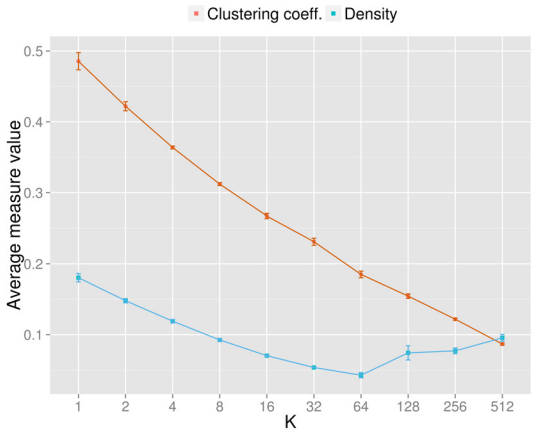
(b) *Total degree*



(c) *Closeness*

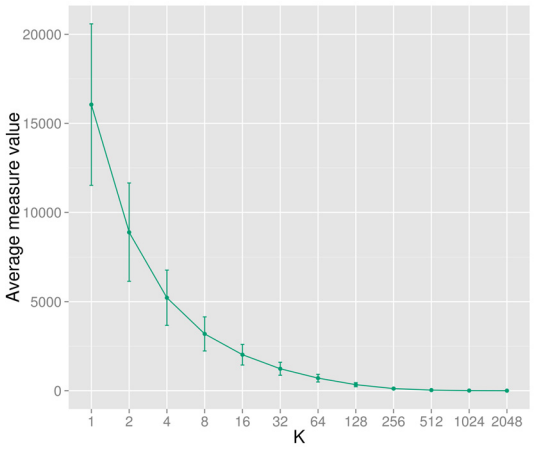


(d) *Number of connected components*

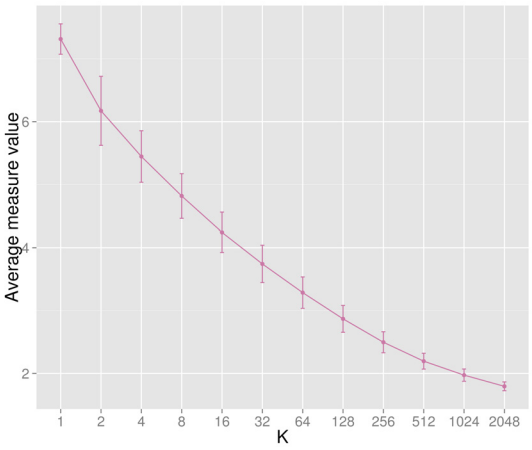


(e) *Clustering coeff. & density*

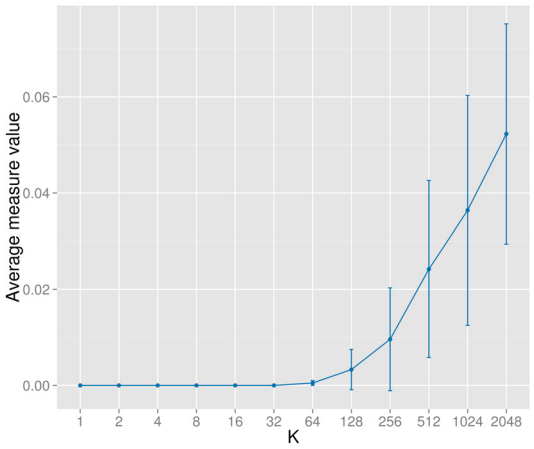
Figure 5.13: Network measures in terms of changing K - Manufacturing company, $z = 1$



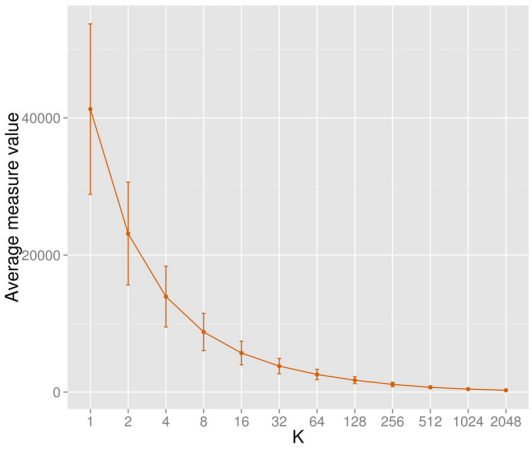
(a) *Betweenness*



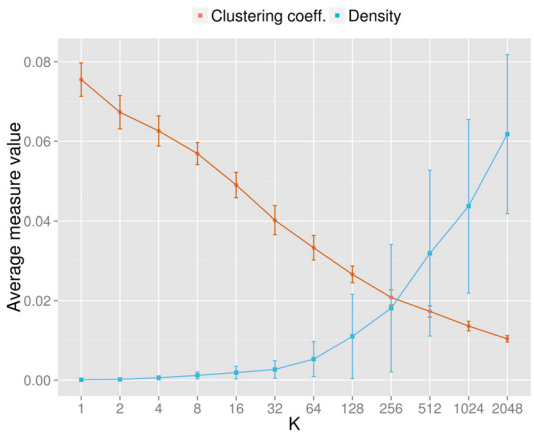
(b) *Total degree*



(c) *Closeness*

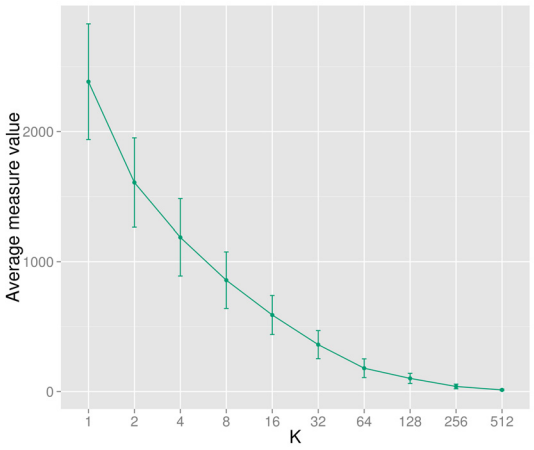


(d) *Number of connected components*

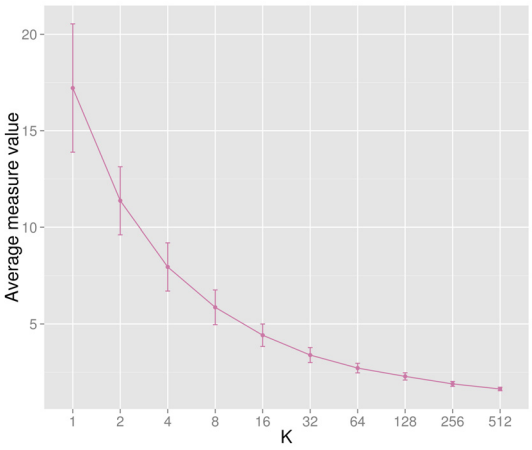


(e) *Clustering coeff. & density*

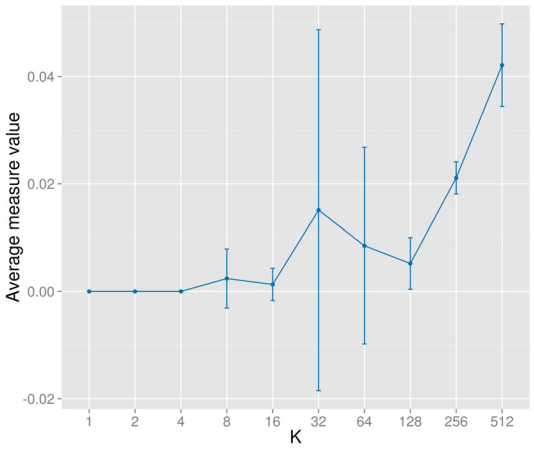
Figure 5.14: Network measures in terms of changing K - Enron, $z = 2$



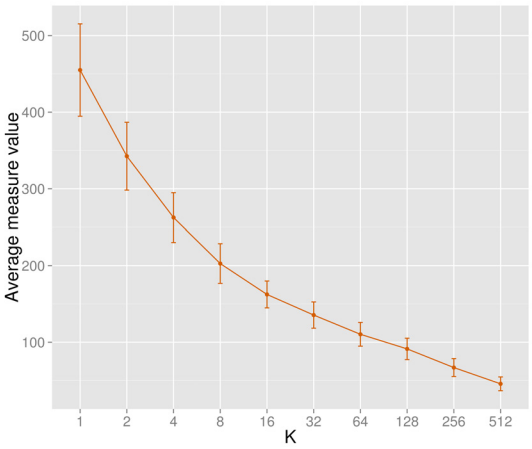
(a) *Betweenness*



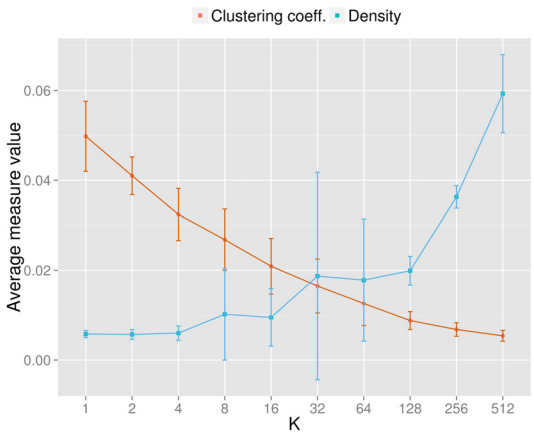
(b) *Total degree*



(c) *Closeness*

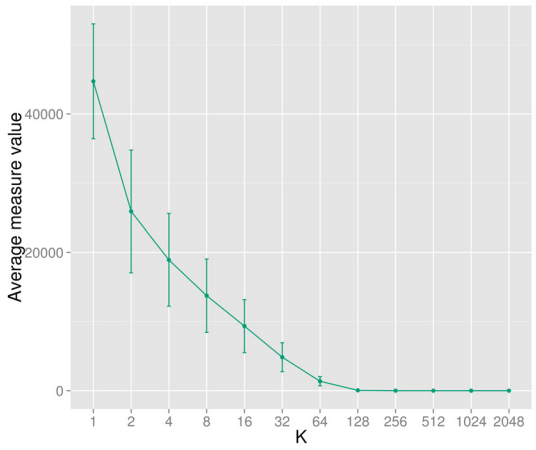


(d) *Number of connected components*

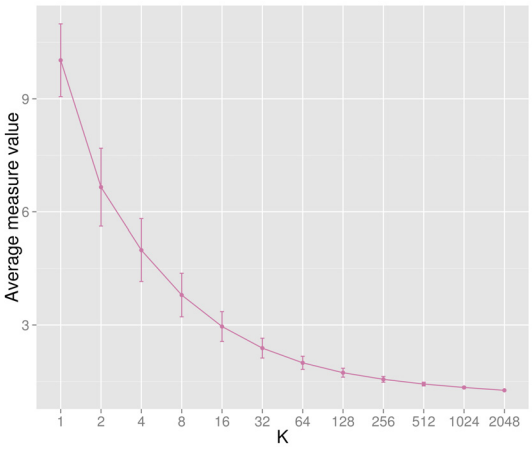


(e) *Clustering coeff. & density*

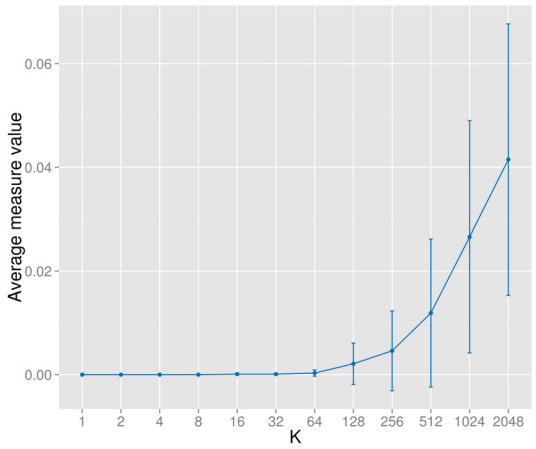
Figure 5.15: Network measures in terms of changing K - University of California, $z = 3$



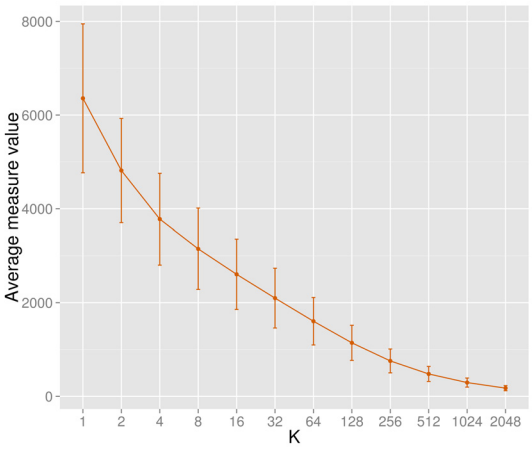
(a) *Betweenness*



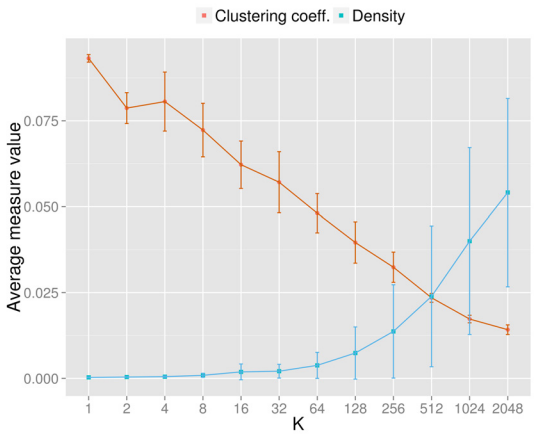
(b) *Total degree*



(c) *Closeness*

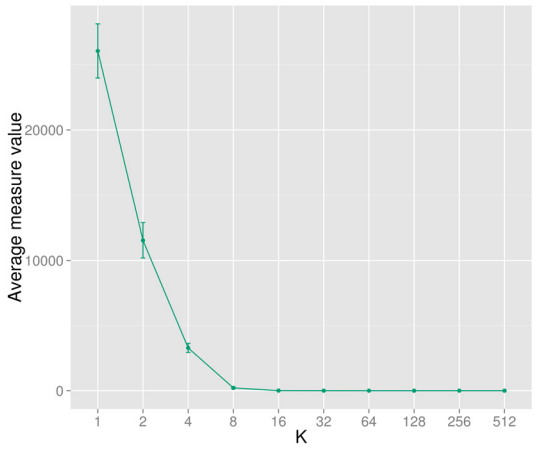


(d) *Number of connected components*

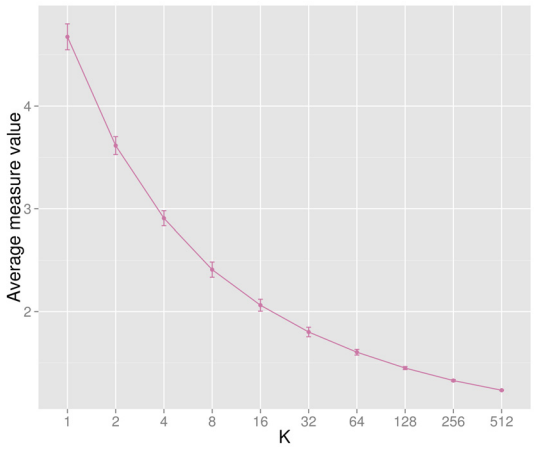


(e) *Clustering coeff. & density*

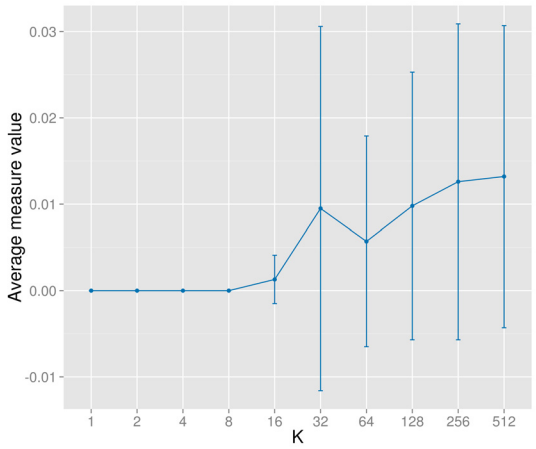
Figure 5.16: Network measures in terms of changing K - Facebook, $z = 4$



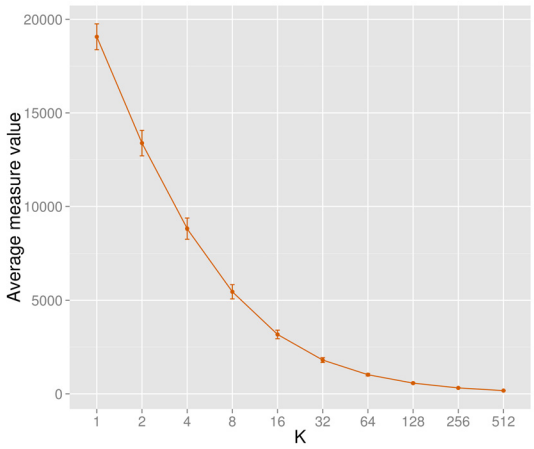
(a) *Betweenness*



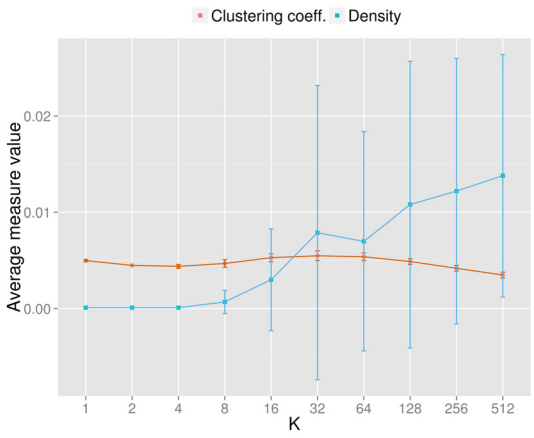
(b) *Total degree*



(c) *Closeness*



(d) *Number of connected components*



(e) *Clustering coeff. & density*

Figure 5.17: Network measures in terms of changing K - Digg, $z = 5$

Table 5.12: The average number of events per individual for TESes for training and testing periods.

Dataset index z	$TES^{Tr}_{D_z,T^{Tr}_u}$		$TES^{Te}_{D_z,T^{Te}_u}$	
	Avg. no. of events per individual	SD	Avg. no. of events per individual	SD
1	101.18	1.37	113.36	3.90
2	4.28	0.28	9.61	1.06
3	8.69	1.89	6.25	0.78
4	5.54	0.94	11.18	1.55
5	0.79	0.02	1.60	0.03

5.3.4.1 Time-limited Event Sequences

Before going into the detail of TSNs analysis, it is worth focusing on the properties of TESes (Table 5.8). They present how different shifts influence the global properties, such as the number of individuals, number of events and the average number of events per individual. In general, this information combined with information provided in Tables 5.1–5.5 and Figures 5.2–5.6 gives a good understanding of the datasets that are used in this thesis. However, since the general properties of datasets were described in Section 5.2.2.6, the discussion here focuses on the TESes itself.

Table 5.8 presents the information for both, the training period as well as the testing period, while the other tables and figures reflect the training period. This decision was taken intentionally, since in the influence maximization challenge for temporal social networks case, knowledge about the future should remain unknown. However, the information about testing periods presented in Table 5.8 gives a brief idea of what to expect and how to interpret future results of spread of influence.

Firstly, comparing the training and the testing periods, it is observed that the most stable TES are the ones that use dataset D_1, namely $TES^{Te}_{D_1,T^{Te}_u}$. The number of individuals is slightly lower than in the training periods, but no new nodes join. The average number of events per individual is 113.36 with standard deviation of 3.90 for the testing period and 101.18 ($SD = 1.37$) for the training

period for all shifts u. TESes built for other datasets introduce bigger differences between the training and testing periods - see Table 5.12. Other TESes have both values different - the number of new individual and the average number of events per individual. An interesting case is for the Enron data - here comparing the training and the testing periods, nearly three quarters of new individuals appear, so the exchange is quite significant and the upcoming task of influencing nodes may become hard, since the majority of nodes join after the training period - nothing is known about them.

Secondly, for TESes built from datasets D_4 and D_5, the number of events in the testing period is increasing, while for others the decreasing trend is observed when moving the shift forward. Moreover, for D_2, D_4 and D_5 the the "productivity" in the testing period expressed as the number of events per individual doubled compared to the training period. The only TESes that show fewer events per individual are those built from the University of California messages - D_3.

So it must be concluded that each dataset introduces its own characteristics that make the TESes built from it somewhat different from others. Moreover, for some TESes the shift introduces bigger changes than for others, revealing that the starting point of the training period is also important, since the network may simply look different. In Chapter 6 how these characteristics influence the spread of influence process will be presented.

5.3.4.2 Temporal Social Networks

Now, having more detail on the TESes, it is worth comparing the TSNs built from those. As it was presented in Table 5.9, there were 54 configurations of TSNs and knowing that each of them was built for different shift values u, it resulted in 270 temporal social networks.

Firstly, interesting conclusions may be drawn from Table 5.11. It shows that the nature of individuals differs from dataset to dataset and from time window to time window. Since some of the conclusions regarding how individuals behave in consecutive time windows have already been presented in Section 5.3.2.1, only a short summary is presented here. In general it is observed that the interchangeability of network nodes depends on the dataset used to build the TSN. However,

when analysing this property at a temporal network level it is observed that for some temporal networks the percentage of similarity is decreasing faster than for others with increasing values of K. It means that individuals in these networks do not exist longer than the length of the period, and they interchange even faster. An interesting phenomenon was also observed when studying the stability of edges. Quite reasonably the similarities between edges in consecutive time windows are smaller than for nodes, since relationships are less stable than individuals in the network. Still, for some networks the stability of edges when comparing consecutive time windows is dramatically small: zero for the Digg dataset or less than 0.15 for University of California. This means that relationships there are very unstable, since they are not transferred to the next time windows. This may support the process of influence, since nodes often change their neighbours. Naturally, it is dependent on weights w and thresholds θ, but the majority of "jumping" nodes may be helpful in terms of overcoming some stalled situations when new nodes cannot be influenced because of a stable network structure.

When comparing the "core" nodes, i.e. nodes that exist in every time window of the TSN (see Figure 5.11), it is observed that with increasing K for most of the network the number of core nodes decreases quickly. This is true for four of five groups of TSNs, only TSNs built from D_1 reveal greater stability of core nodes. This is because when comparing these TSNs to others the dataset they were built from is the most stable and non-changing one.

Lastly, the analysis of structural measures of TSNs shows that some interesting phenomena can be observed (see Figures 5.13–5.17). As betweenness and total degree measures behave as expected (except for Figure 5.13a where the betweenness rises), the number of connected components decreases. This goes against intuition, since if the granularity increases, there should be more connected components, since contacts become sparse. This actually means that with increasing K the most stable relationships and nodes stay and the temporal components vanish.

However, the most interesting part of this analysis is presented in Figures 5.13e–5.17e. It is observed that at some point the values of density start to rise exponentially. It means that at some point the network structure of component SNs becomes more complete.

5.4 Summary

This chapter presented what operations were performed to move from the level of datasets through time-limited event sequences to, finally, temporal social networks. For each of these entities their properties were presented and discussed. Moreover, the framework for conducting the experiments was introduced and briefly described.

Finally, the resulting TSNs were analysed in terms of network properties and measures to see how they change if granularity differs and some interesting points were discussed.

It is worth noticing that the operations performed in this chapter were very high-level, i.e. instead of looking at a local phenomenon that may be observed in particular social networks, a global perspective is used. The temporal social networks are compared by averaging the values of particular measures at a social network level. This is rather a non-standard approach, but as it will be presented in the next chapter, it may provide satisfying results.

Chapter 6

The Method for Maximising the Spread of Influence in Temporal Social Networks

6.1 Introduction

In this chapter a new method for maximizing the spread of influence is presented - *tInf*. The name comes from *t*emporal *Inf*luence and suggests that this approach benefits from the temporal properties of the social networks. Indeed, to ensure good results within reasonable time frame, this method takes advantage of what has been observed in the previous chapter. Namely, when generating temporal social networks with a constantly increasing number of non-overlapping time windows of equal length, at some point the properties of TSNs will change in a desired way allowing to choose well-suited seeds. As it is presented later in this chapter, those seeds used for the process of social influence may greatly improve the results.

This chapter is organized as follows. Firstly, the problem of finding seeds in social networks is quantified to show its complexity. Before the proposed method will be introduced, a number of arguments are provided as to why the temporal approach is a must when studying the social influence processes in social networks. The main reason is that the influence process behaves differently for

time-aggregated networks and for temporal social networks. This is why in Section 6.3 a new model for spread of influence in temporal networks is introduced. Section 6.4 a new concept of seed selection in temporal social networks is proposed and discussed. In Section 6.5 the comparison of static and temporal approaches is presented in order to show why it is reasonable to follow the temporal path when analysing the process of spread of influence.

Next, in Section 6.6 the proposed *tInf* method is introduced and described. In Section 6.7 the experimental setup is described. Experimental results are presented in Section 6.8 allowing the evaluation of the proposed method, and to see when it provides the best outcomes. The experiments start with showing important aspects of social influence, such as the role of seed set size or the threshold level in the process, since those are also important research questions that should be addressed before thinking about implementing seed strategies. Then the proposed method is studied in terms of results, performance and it is also compared to the greedy algorithm. Lastly, Section 6.10 concludes this chapter.

6.2 Motivation: Quantifying The Challenge

As it is presented in Section 4.3, finding the optimal solution for the time-aggregated social networks is a NP-hard problem. For the most popular definition of the problem (Kempe et al. [2003]), the optimal solution is considered as the seed set, seeds in short, that maximize the spread of influence for a given influence propagation model within a given budget c. The budget is most often considered simply as number initially influenced, so that $|\Phi(0)| = c$ and the cost for influencing an individual is also fixed for all the nodes. Because of the hardness of the problem, many heuristics are proposed to obtain acceptable results - the development of the solutions is presented in Section 4.4. However, at this point, it is actually worth showing the real numbers for the real world social networks in order to present how big is the search space when looking for optimal solution.

Assuming that the budget c represents the initial percentage of nodes that will be influenced from a given set of nodes that appeared in the social network

to the moment of influencing nodes, the number of combinations is expressed as:

$$C^{c}_{|V|} = \binom{|V|}{c} = \frac{|V|!}{c!(|V|-c)!}, \tag{6.1}$$

where $|V|$ denotes the number of nodes that exist in the social network. In the temporal variant of the problem it is the number of nodes that exist before the influence process starts, so it is the set of nodes in the training TES (see Figure 5.7).

In this thesis two values of c are evaluated. Here, the budget is not expressed as a number of nodes, but as a percentage of nodes from the whole set of nodes that can be influenced initially. The budged is then $c = 0.05$ for the smallest dataset D_1 and $c = 0.01$ for other datasets (rounded up). Despite that the experimental setup is described later in Section 6.7, these values are now used to show how many combinations have to be evaluated in order to numerically find the optimal solution. Table 6.1 presents the complexity of this task for all the evaluated TES^{Tr}, since the set of unique nodes in TSN and TES is the same. This table shows that the numerical approach of evaluating all possible combinations is impossible and gives an argument why the heuristics are needed.

6.3 The Idea and Importance of the Temporal Approach

Another important research question is whether the temporal approach is really needed and whether the time-aggregated social networks are not enough to model the dynamic processes of spread of influence. In fact, the question is whether by using the static social networks the paths of the process and its final outcome is similar, especially when thinking about maximizing the spread of influence.

6.3.1 Static Approach

In this thesis the Linear Threshold model LT is considered (see Section 3.4.3.2 for its description). It introduces the threshold parameter θ_{v_i}, which represents the weighted fraction of neighbours of node v_i that have to be influenced to make

Table 6.1: Number of combinations for choosing seeds in defined TESes.

Dataset index z	Shift index u	$TES^{Tr}_{D_z,T^{Tr}_u}$			
		No. of individuals	Budget c	No. of seeds	Number of combinations
1	1	166	0.05	9	2.115317×10^{14}
1	2	165	0.05	9	2.000631×10^{14}
1	3	157	0.05	8	7.637643×10^{12}
1	4	156	0.05	8	7.248464×10^{12}
1	5	156	0.05	8	7.248464×10^{12}
2	1	30,331	0.01	304	uncountable
2	2	36,130	0.01	362	uncountable
2	3	47,165	0.01	472	uncountable
2	4	54,272	0.01	543	uncountable
2	5	64,629	0.01	647	uncountable
3	1	1,786	0.01	18	4.900029×10^{42}
3	2	1,759	0.01	18	3.719998×10^{42}
3	3	1,707	0.01	18	2.161735×10^{42}
3	4	1,542	0.01	16	4.51668×10^{37}
3	5	1,389	0.01	14	1.068754×10^{33}
4	1	12,045	0.01	121	uncountable
4	2	14,323	0.01	144	uncountable
4	3	17,150	0.01	172	uncountable
4	4	19,904	0.01	200	uncountable
4	5	22,476	0.01	225	uncountable
5	1	21,132	0.01	212	uncountable
5	2	22,638	0.01	227	uncountable
5	3	23,437	0.01	235	uncountable
5	4	23,594	0.01	236	uncountable
5	5	23,608	0.01	237	uncountable

this node influenced as well. Intuitively, comparing time-aggregated and temporal social networks as defined in this thesis (see Definition 4), the static networks will be more complete, i.e. nodes will have a higher degree, since the edges become aggregated and for the temporal case they appear and disappear resulting with a lower average degree. If the same number of seeds is selected, for the static case it might be a problem with finding appropriate seeds that will start a cascade of influence in the social networks. To give an example, if in a temporal social network an average in-degree of a node is four, $\theta = 0.50$ is constant for all nodes and weights are assigned as a fraction of one and number of neighbours of a given node, on average two neighbours are required to influence the node. In the static case the average degree will rise, therefore the number of influenced neighbours needed will be higher and the budget c is still the same. However, as it was already stated in Section 2.6, the dynamics are embedded in social networks and trying to consider whether better results will be obtained for a static social network or a temporal one is pointless knowing that networks cannot be forced to become static. The only question is how different the outcome of the process will be assuming that the networks are static.

What a static approach looks like is presented in Figure 6.1. Here, a static (time-aggregated) social network is used both for seeding and evaluation. Since the network does not change, seeding may be performed at any time. As the LT model assumes that the stop condition is met when no more nodes can be influenced, it is not necessary to define the number of iterations at the beginning. In this figure it is assumed that the number of iterations after which no more nodes could be influenced was K' for illustration purposes - this number is not defined *a priori*. The spread of influence SI (see Definition 5) is then the total number of influenced nodes when the stop condition is met.

6.3.2 The Model of Temporal Spread of Influence

The temporal model for spread of influence defined in this section is intended to provide an environment in which it will be possible to evaluate different seeding strategies for temporal social networks. Not to mention about anything related to the process of seed selection methods, it focuses only on the evaluation part.

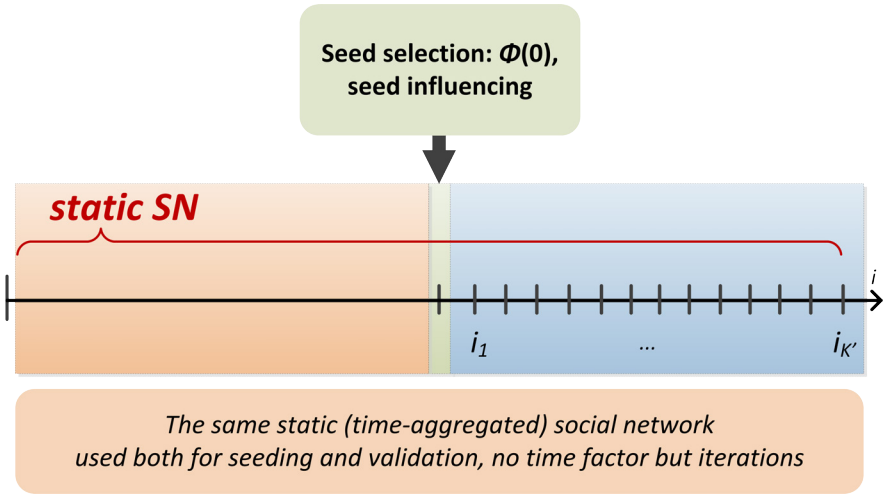


Figure 6.1: The static approach in the spread of influence.

Before it will be described and discussed, it is presented in Figure 6.2.

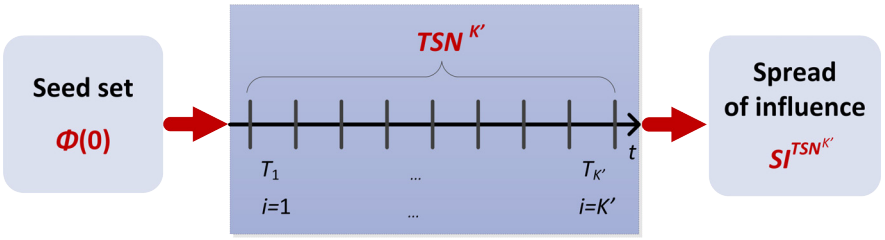


Figure 6.2: The temporal model for spread of influence.

This model introduces the temporal social network $TSN^{K'}$, which consists of K' social networks (see Definition 4 for details). In the beginning a number of seeds is being influenced, $\Phi(0)$. Then, in K' iterations the influence spreads in the temporal social network $TSN^{K'}$ where in each time window, a particular social network SN is the underlying layer for influencing nodes. So, starting with the first social network built over period T_1, the influence process takes place there (a single iteration) and then the next iteration takes place in SN built over period T_2 and so on. At the end, after the time window $T_{K'}$, spread of influence $SI^{TSN^{K'}}$ is calculated as the number of influenced nodes after $T_{K'}$ time windows, i.e. $|\Phi(T_{K'})|$, see Definition 6.

The most important difference between the proposed model and the static approach presented in Section 6.3.1 is that instead of the static social network, the

temporal social network is being used in the process of the spread of influence. Moreover, instead of the natural stop condition typical of the static network, where the process stops when no more nodes can be influenced, this model assumes that the value K' is given. It is the stop condition that is needed in order to evaluate different seeding strategies. Naturally, the $TSN^{K'}$ is built over a chosen TES, see Definition 3. In this case this period covered by TES is being split into K' time windows of equal length forming the TSN.

Please note, that this model is does not take into account the method of seed selection and its purpose is only the evaluation of seed selection strategies for temporal social networks. Seeding may be of any kind, e.g. random or using some seeding strategy, and performed on any type of layer: time-limited event sequence, social network, or temporal social network. However, in Section 6.4 a concept of seed selection in temporal social network is introduced and discussed.

In order to verify how the static approach and the temporal model differ in terms of spread of influence, in the Section 6.5 there is an experimental comparison of the two approaches presented. But before this comparison will be performed, it is necessary to introduce a concept of seed selection in temporal social networks.

6.4 A Concept of Seed Selection in Temporal Social Networks

Having the model of temporal spread of influence defined, one of the most important problems now is to choose the winning strategy for selecting seeds. As it is presented in Figure 6.2, this model does not cover the process of choosing seeds. It assumes that the seed $\Phi(0)$ is given, and evaluates the quality of seeds by using a temporal social network $TSN^{K'}$. At the end the total number of influenced is provided - $SI^{TSN^{K'}}$. At this point consideration switches to the method of how to choose seeds in order to maximize the spread of influence under the so defined model.

Firstly, it has to be assumed, that while the spread of influence takes place in the unknown future, the past is known. this means that the researcher has access to the past events that occurred between individuals. As it was presented

in Chapter 2 there is a number of ways this past may be expressed: as an event sequence, a social network or a temporal social network. Here, it is assumed that the most detailed information is available, namely the event sequence, see Figure 6.4. On the left-hand side of this figure the training set is presented, while on the right-hand side the testing part is observed and the testing part follows the model of the temporal spread of influence (Section 6.3.2).

As the seed set has to be chosen in the training part, only the information about nodes that appeared in the past is known and $\Phi(0)$ will contain only a subset of these nodes, having the size c (budget). Naturally, it is expected that most of these nodes from the seed set will appear in the future in order to not to waste the budget on seeds without technical ability to influence other nodes, yet, it is not guaranteed. One of the strategies might be to focus on nodes that appear recently, since their probability of appearing soon is expected to be higher than those that appeared at the beginning of the training period. This leads to the conclusion that the time factor might play a crucial role for seeding strategies. Unlike the greedy method (see Section 4.4.3) that is suited for static networks, the temporal information may be the key to a better quality of seeding strategies. Naturally, the greedy method or other methods may also be used, since the TES may be converted into the time-aggregated social network. Yet, the question is whether it will outperform the time-respecting methods for a so-defined model of temporal spread of influence.

To conclude the above short discussion, the basic taxonomy of seed selection methods considers time-respecting methods and those that do not use the time factor. For time-respecting methods, it is possible to build seeding strategies using either event sequence, time-limited event sequence or temporal social network. For static methods, only building a static social network is possible, however it may also be built using the whole period of the past events or a limited one. This leads to another important question - how far should we look into the past? Is it reasonable to use the whole events history or narrow down the events' period to these recent ones? Is there any difference if we will shorten the training period for seeding - will the results of the spread of influence remain the same? These questions indicate that the problem of spread of influence for temporal social networks definitely becomes more complicated than for static networks. Yet, it

has to be faced, since most of the surrounding networks are dynamic.

Another interesting challenge may also be introduced at this point which is related to the problem. If the problem is stated differently, i.e. we want to maximize the spread of influence at time T'_K, until what time should we observe the network to obtain enough data in order to apply the seeding strategy knowing we have no access to the past data. In this case it is required to find the trade-off between the period for training and the time for the spread of influence process. This problem has not been challenged yet at all.

Still, coming back to the problem of seeding in temporal social networks, it is assumed that there is the need to benefit from the past temporal data. As it has been stated above, it is possible due to the use of event sequence that may be used as is or that can be transformed into the temporal social network. The concept of seed selection in temporal social networks as defined in this thesis is using the latter approach, retaining the chance to perform classical SNA for component SNs forming the TSN. This concept is defined as follows.

In order to perform the seed selection in temporal social networks, transform the available ES/TES covering the past to TSN consisting of K windows. The number of windows does not need to be the same as K' used in the model for temporal spread of influence (see Section 6.3.2). Then, apply a seeding strategy that respects the time factor. A number of seeding strategies that implement this approach is presented in Section 6.6.5.

The proposed concept of seed selection for temporal social networks combined with the model of temporal spread of influence is presented in Figure 6.3.

The proposed concept will be then compared with two static approaches, namely the greedy method and the random one. However, in order to fully benefit from the temporal information a method for maximizing the spread of influence in the temporal environment is proposed - *tInf*. Its goal is to maximize the spread of influence by taking the advantage of temporal information that is available. Yet, just before, to provide an argumentation why the temporal approach is a must, a comparison of static and temporal approaches is presented in the next section.

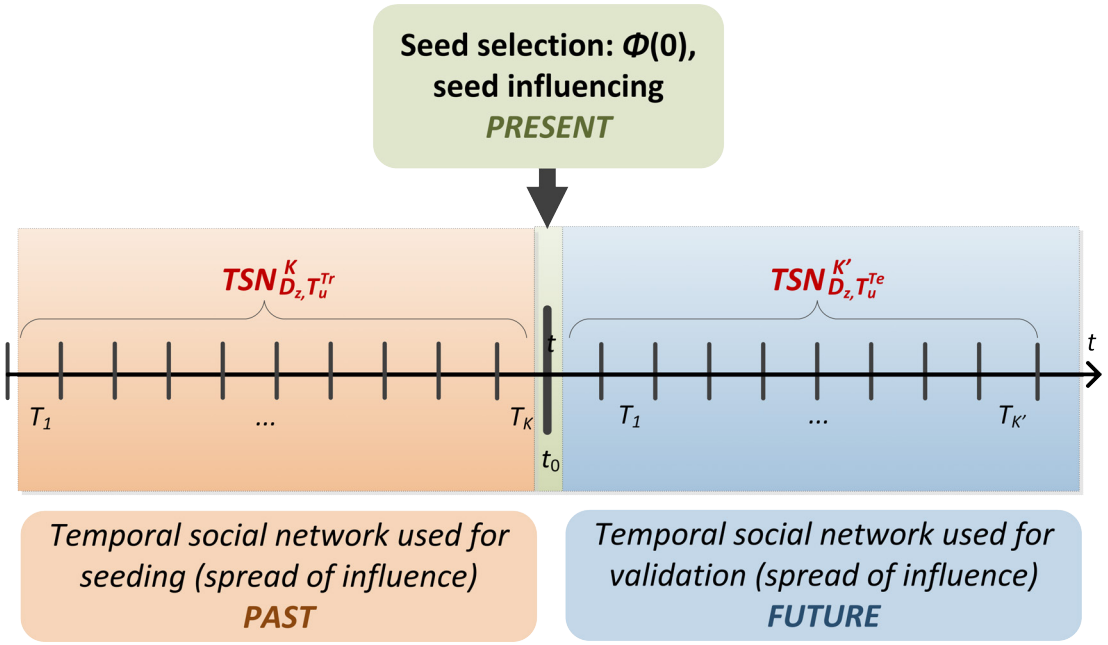


Figure 6.3: The concept of seeding in temporal social networks.

6.5 Comparing the Static and Temporal Approach

In order to verify what is the difference between running the influence process in a static and temporal environment, a small experimental scenario is evaluated for two different datasets - D_1 (manufacturing company) and D_3 (University of California). The method for choosing seeds is as follows. For a given budget find c percent of nodes in the training event sequence $TES^{Tr}_{D_z,T^{Tr}_u}$ that initiate the highest number of events. This method is relatively simple and does not require that a static social network nor a temporal social network is built. Simply, the nodes that initiated the highest number of events are selected. These nodes form the initial seed set $\Phi(0)$, see Section 4.3.2 for details.

After choosing seeds and influencing them, the evaluation takes place for the model of temporal spread of influence with $K' = 10$ (see Section 6.3.2). For the first dataset the first influence process occurs in $TSN^{10}_{D_1,T^{Te}_1}$ and the for second one in $TSN^{10}_{D_3,T^{Te}_1}$. The propagation model is the LT model with a fixed threshold $\theta = 0.50$ for all nodes. The budget c is 0.05 for $TES^{Te}_{D_1,T^{Te}_1}$ and 0.01 for $TES^{Te}_{D_3,T^{Te}_1}$. The details of the experiment are summarized in Table 6.2 and the approach is

presented in Figure 6.4.

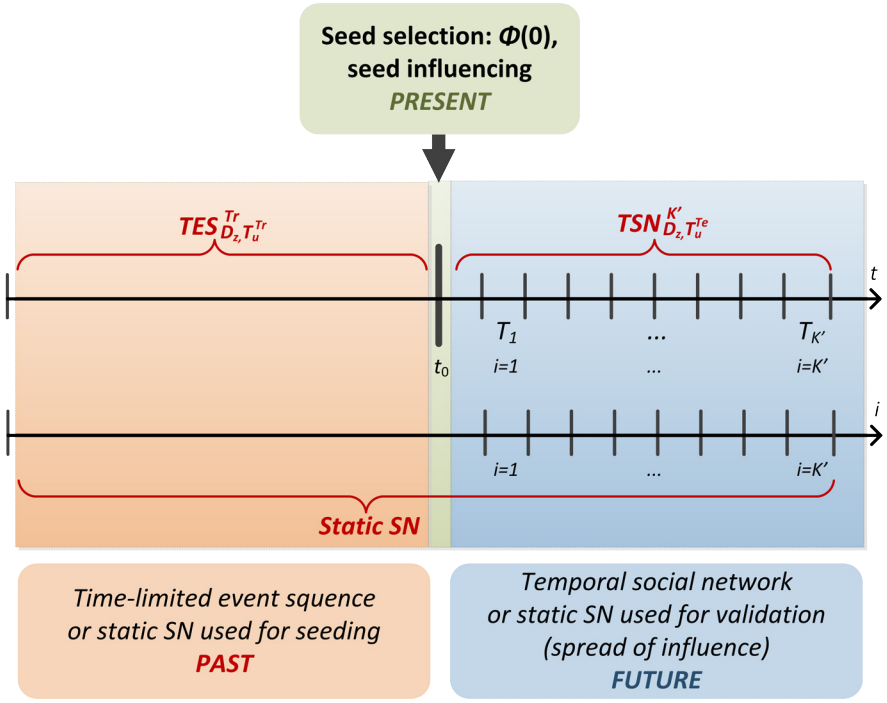
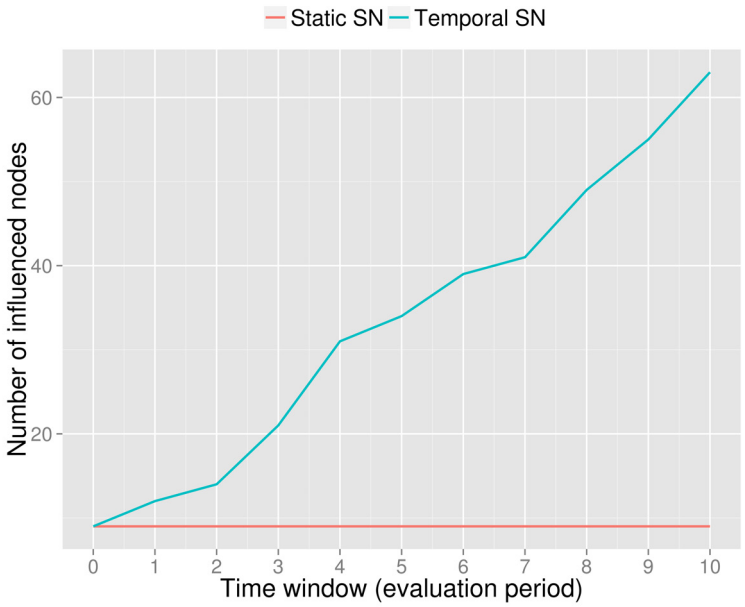


Figure 6.4: Comparing the process of spread of influence for the temporal (top) and static approaches (bottom).

Results of this process are presented in Figure 6.5. It is clearly observed that these processes do not follow each other for the static and temporal cases. In Figure 6.5a for the time-aggregated network no new nodes were influenced apart from the initial seed set, while for the temporal case 63 nodes were activated. For the University of California networks (Figure 6.5b), in both scenarios the influence spread was an incremental process and for the static social network it end up with sixty nine influenced nodes, while for the temporal approach the outcome was three hundred and five nodes. It is not necessarily true that the outcome for the process will be smaller for the time-aggregated networks. For instance, for $\theta = 0.33$ for the University of California dataset, the static case outperforms the temporal one by about 26%. Still it should be remembered that it is impossible to force the network to be static, since this process is outside the control of the researcher. This is why the only possible direction of future research in the area of dynamic processes are temporal networks.



(a) *Manufacturing company*

(b) *University of California*

Figure 6.5: Comparing the outcome for spread of influence for static and temporal social networks.

Table 6.2: Parameters for the spread of influence process.

Parameter	$TES^{Te}_{D_1,T_1^{Te}}$	$TES^{Te}_{D_3,T_1^{Te}}$
Propagation model	Linear Threshold	Linear Threshold
Budget c	0.05	0.01
Number of nodes (training)	166	1,786
Number of seeds	9	18
Threshold θ	0.50	0.50
Number of windows (testing) K'	10	

6.6 A Method for Maximizing the Spread of Influence - *tInf*

6.6.1 Introduction

In Chapter 5 the properties of experimental temporal social networks are presented. Those networks are generated by using five different datasets. Those temporal networks have a different resolution, varying from $K = 1$ to $K = 2048$, i.e. they consist of 1 to 2^{11} social networks. The temporal social network consisting of one social network is, naturally, a time-aggregated one. As it is observed, the structural properties of the component social networks vary depending on the value of K and this is the basic idea behind the proposed *tInf* method.

When studying the phenomena of temporal social networks, it was observed that the varying resolution of TSNs is important when applying seeding strategies. In Michalski et al. [2014] it is shown that the greater the resolution of temporal network, the better the results of spread of influence. In fact, this work is preliminary research into what is presented in this thesis, since only a couple of temporal social networks were compared, consisting of a single time window, five and ten time windows. In this work it was expected that this process of increasing the spread of influence by using more and more time windows has its limits, but this was not studied further, as well as no analysis being made as to why this phenomenon appears.

In this thesis, the extensively extended approach is presented which is named *tInf*, from *t*emporal *Inf*luence. This method starts with creating a number of temporal social networks from event sequences or time-limited event sequences.

Next, in order to benefit from temporal network properties, a number of methods are introduced which reflect the timed aspects of nodes (see Section 6.6.5) and they serve as seed selection strategies. The purpose of this activity is to find out whether there is any correlation between the network structural properties and the spread of influence and whether there is any chance of finding a temporal social network resolution that will allow maximization of the spread of influence for a given propagation model. The model chosen in this thesis is the Linear Threshold, see Section 3.4.3.2 for details on it. The *tInf* method is described in detail in the following section.

6.6.2 The Concept of the Method

The goal of the *tInf* method is to maximize the spread of influence by avoiding a computationally complex search through the solutions space. As it is presented in Section 6.2, the task of numerically finding the optimal solution for the spread of influence in time-aggregated networks is almost impossible and the analytical solution complexity is NP-hard. Moreover, as it is presented in Sections 4.3 and 4.5 the problem for temporal social networks is yet to be properly addressed and defined, since no commonly agreed definition of it exists so far.

In this work the problem is defined as in Section 4.3.2. That is, for a certain temporal social network TSN^K it is required to find a certain set of nodes limited by budget c that will be influenced. Those nodes are denoted as seed set $\Phi(0)$. In consecutive time windows $T_1, \cdots, T_{K'}$, representing the future, those seeds influence other nodes that also become influencers. Finally, after the K'th time window, the overall number of influenced nodes ($SI^{TSN^{K'}}$) is expected to be the highest. As it is presented in Sections 5.2.4 and 6.6.1, the training TSN used for choosing seeds is denoted as $TSN^K_{D_z,T_u^{Tr}}$, while the one used for evaluation based on the testing period is $TSN^{K'}_{D_z,T_u^{Te}}$, where K and K' denote the number of windows of which particular TSN consists, in training and testing sets, respectively. D_z represents the dataset used and T_u^{Tr} and T_u^{Te} denote the period of time-limited event sequence for which the TSNs were constructed - training and testing, respectively. This process is presented in Figure 6.3. The left-hand side of this figure represents the $TSN^K_{D_z,T_u^{Tr}}$, while the right-hand side - $TSN^{K'}_{D_z,T_u^{Te}}$.

In this thesis K' for the testing TSN equals ten.

Algorithm 4 Algorithm *tInf* for maximizing the spread of influence SI in temporal social networks

1: input $TES^{Tr}_{D_z,T^{Tr}_u}$, K^{set}, budget c
2: initialize $\Phi(0) = \emptyset$, $\Psi = \emptyset$, $\Omega = \emptyset$, $\Omega^{sort} = \emptyset$
3: **for** $K \in K^{set}$ **do**
4: generate $TSN^K_{D_z,T^{Tr}_u}$
5: $\Psi = \Psi \cup TSN^K_{D_z,T^{Tr}_u}$
6: **end for**
7: $p = FindBestK(K^{set}, \Psi)$
8: $\Omega = SM(TSN^K_{D_z,T^{Tr}_u})$
9: $\Omega^{sort} = order(\Omega)$
10: **for** $i = 1$ to $ceil(c * |\{V_1, \ldots, V_p\}|)$ **do**
11: $\Phi(0) = \Phi(0) \cup vert(\Omega^{sort}(i))$
12: **end for**
13: output $\Phi(0)$

The proposed method *tInf* is presented as Algorithm 4. It requires $TES^{Tr}_{D_z,T^{Tr}_u}$ - the time-limited event sequence, and K^{set} - the set of values of K to be evaluated.

Firstly, this algorithm initializes an empty set $\Phi(0)$ representing the final seed set. Next, the *tInf* algorithm finds the value p, such that for all evaluated values of $K \in K^{set}$ a particular condition is met, i.e. the best K from K^{set} has been found.

If p is obtained, for the temporal social network TSN^K a chosen measure SM is calculated which returns the set of tuples $(v_i, sm(v_i))$, i.e. values of this measure SM for each node. Next, Ω is ordered in a given way by the seeding strategy and Ω^{sort} becomes a list of ordered elements of set Ω, see Section 6.6.5.

Lastly, the budget c limits how many nodes nodes from the list Ω^{sort} should be added to the seed set $\Phi(0)$. The function $vert(\Omega^{sort}(i))$ obtains the node from the Ω^{sort}. The evaluation which is outside of the method itself takes place by using $TSN^{K'}_{D_z,T^{Te}_u}$. The *tInf* method altogether with the validation process is presented in Figure 6.6.

The key part of this algorithm is to find the appropriate K value, i.e. line 7 of Algorithm 4. This will be discussed later, but assumes that there exists a method of finding the best K of the K^{set}.

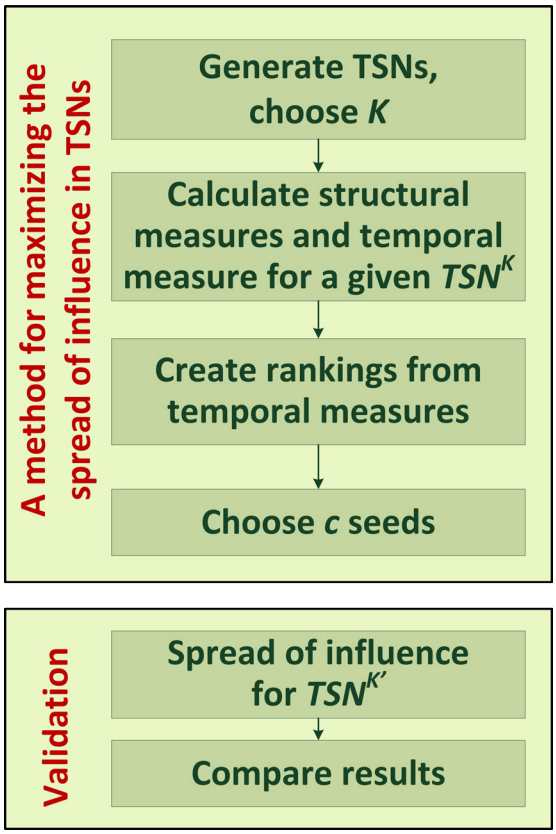


Figure 6.6: The *tInf* method for maximizing the spread of influence in temporal social networks.

The *tInf* algorithm uses a measure SM which is calculated for constructing the rank of nodes Ω and choosing seeds. It may be any seeding strategy, such as nodes with top or bottom in-degree, highest betweenness etc. and in this thesis various methods for node rankings were considered. All of them were based on different structural network measures but for temporal networks, diverse forgetting methods were applied to respect the new knowledge more than the old, see Section 6.6.5.

Note that the temporal approach provides a unique chance to utilize the dynamics of the social network observed in the past. If these networks' dynamics (kinds and speed of changes) are to some extent similar in the future, the time-sensitive seeding may potentially deliver better results. The novelty of the method is that it attempts to find the best K for applying the time-sensitive measures, i.e. it assumes that by finding this value of K the time-sensitive seeding strategies will perform best.

6.6.3 The Dependency between the Spread and Structural Features

The proposed algorithm *tInf* assumes that it is possible to find the best value of K, i.e. the best granularity of the temporal social network. For this granularity the seeding strategy applied should perform the best in the testing period. Naturally, one way of finding it in a given scenario is just to perform all evaluations for all the TSNs generated for K^{set}. Yet, this is approach is useless, since it is using *a posteriori* knowledge about the evaluation period. This is why it is necessary to propose a method that will provide good results without comparing the process outcomes. In order to achieve that goal, the function $FindBestK$ from Algorithm 4 (line 7) is proposed to be a linear regression task (Neter et al. [1996]). Here, the estimated parameters of the linear model represent the weights for each input measure. The formal representation of the solution is presented in Equation 6.2.

$$SI_i = a_1 * m_{i,1} + a_2 * m_{i,2} + \ldots + a_j * m_{i,j} + \epsilon, \quad i \in \{1, \ldots, |TSN^{set}|\} \tag{6.2}$$

In the above equation SI_i represents the spread of influence for a given temporal social network (see Definition 6), $a_1, \ldots, a_j$ are the parameters of the model and $m_1, \ldots, m_j$ represent particular regressors (input variables). The TSN^{set} represents the set of evaluated TSNs, i.e. all the TSNs that were used as for learning. In this thesis the input variables are the following:

- $m_1 - AvgBetweenness$
- $m_2 - AvgTotalDegree$
- $m_3 - AvgCloseness$
- $m_4 - AvgClustering$
- $m_5 - AvgDensity$
- $m_6 - AvgNoOfConnectedComponents$

These are the same averaged measures as evaluated in Section 5.3.3.3. However, the measure values and the spread of influence SI were normalized linearly according to their min-max value, so that they are in the range $[0, 1]$, 0 corresponds to the minimum value and 1 to the maximum value. The idea is to find the best value of K by performing linear regression for all the evaluated scenarios. Then, knowing the values of the parameters it could be possible to omit the computation of spread of influence for all the configurations, but to apply the linear regression parameters to new data in order to find the best K. Then, for a given K a particular seeding strategy may be applied, as presented in the next section. However, in this thesis it is assumed that the regression parameters are specific for a given dataset, i.e. the regression was performed separately for all the datasets. Due to the fact that the experiments were performed for 5 different shifts u for each dataset, in the evaluation scenario, when performing the regression, the cross-validation method has been used (Picard and Cook [1984]). In this scenario, the TSNs of four shifts were used for obtaining the parameters of the regression model, while the fifth one was used for evaluation - this was repeated five times for different evaluation shifts. Naturally, in the real world scenario, the evaluation would be performed in the unknown future. Yet, to see how well

the proposed method performs, such a cross-validation scenario has been applied. Results are accompanied by the root-mean-square error in order to show how well the linear regression performed.

6.6.4 Finding the best K as a Classification Task

Another approach that may be used in order to determine the best value of granularity K among all the evaluated values in K^{set} could be the classification task (Bishop et al. [2006]). Here, in comparison to the linear regression presented in the previous section, instead of finding the coefficients for every global structural measure, the challenge is defined as follows. Having the results for the greedy algorithm and for other seeding strategies, a binary value is assigned for each K that denotes whether for this particular K a seeding strategy has beaten the greedy algorithm in terms of the spread of influence SI. Then, a classification model is learned, e.g. random forest (Breiman [2001]). Its features are the same values provided as in the Section 6.6.3. Having the trained model, it is possible to test every K in order to check whether for such K better results than for the greedy method can be obtained. However, in this dissertation this approach is given as another example of how to find the best K. The further experimental studies are limited only to the linear regression approach.

6.6.5 Time-sensitive Seeding Strategies

Firstly, three simple seeding strategies were introduced, which allow the ordering of nodes based on calculated structural measures SM (total degree, in-degree, out-degree, betweenness, closeness) respecting all periods in the temporal social network in the accumulated way. If $SM_p^M(v_i)$ denotes the value of a given structural measure M (e.g in-degree) of particular node v_i in the pth time window ($1 \leqslant p \leqslant K$), several unnormalized aggregated measures respecting the temporal aspects of the node's activity in all consecutive periods can be defined as follows and used as ordering strategies in Algorithm 4 (line 9):

- Maximum (Max)

$$Max(v_i) = \max(\bigcup_{p=1}^{K} SM_p^M(v_i)) \tag{6.3}$$

- Minimum (Min)

$$Min(v_i) = \min(\bigcup_{p=1}^{K} SM_p^M(v_i)) \tag{6.4}$$

- Sum (Sum)

$$Sum(v_i) = \sum_{p=1}^{K} SM_p^M(v_i) \tag{6.5}$$

The above aggregations, however, do not make use of the sequential nature of time and general phenomena that recent social relationships are likely to be more influential than old ones. Hence, nine new aggregations that also take into account the "forgetting" aspect of time are introduced, i.e. the value of a given structural measure in the most recent time window is the most important, while the measure's value in the oldest period is the least valuable. The purpose of this, was not only to capture the dynamics of user behaviour but also to emphasize users' latest activities. These new aggregations are defined in the following way:

- Maximum Logarithm ($LogMax$)

$$MaxLog(v_i) = \max(\bigcup_{p=1}^{K} log_{K-p+1} SM_p^M(v_i)) \tag{6.6}$$

- Minimum Logarithm ($LogMin$)

$$MinLog(v_i) = \min(\bigcup_{p=1}^{K} log_{K-p+1} SM_p^M(v_i)) \tag{6.7}$$

- Sum of Logarithms ($LogSum$)

$$SumLog(v_i) = \sum_{p=1}^{K} log_{K-p+1} SM_p^M(v_i) \tag{6.8}$$

- Maximum Power ($PowMax$)

$$MaxPow(v_i) = \max(\bigcup_{p=1}^{K}(SM_p^M(v_i))^p) \tag{6.9}$$

- Minimum Power ($PowMin$)

$$MinPow(v_i) = \min(\bigcup_{p=1}^{K}(SM_p^M(v_i))^p) \tag{6.10}$$

- Sum of Powers ($PowSum$)

$$SumPow(v_i) = \sum_{p=1}^{K}(SM_p^M(v_i))^p \tag{6.11}$$

- Linear Forgetting (Lin)

$$LF(v_i) = \sum_{p=1}^{K} pSM_p^M(v_i) \tag{6.12}$$

- Hyperbolic Forgetting (Hyp)

$$HF(v_i) = \sum_{p=1}^{K} \frac{1}{K-p+1} SM_p^M(v_i) \tag{6.13}$$

- Exponential Forgetting (Exp)

$$EF(v_i) = \sum_{p=1}^{K} \frac{1}{\exp(p)} SM_p^M(v_i) \tag{6.14}$$

All the aggregations combined with all typical node structural measures were used to create node rankings and select the seed set for spreading the influence - they form seeding strategies. The structural measures are the following: in-degree (In), out-degree (Out), total degree (Tot), betweenness (Bet) and closeness (Clo). Those measures are described in Section 5.3.3.2. However, only few of them really provided reasonable and distinct results, which are shown in Section 6.8.3.

Since the above aggregations are combined with network structural measures, the following convention is used to build the name of the evaluated seeding strategy: the abbreviation of the measure and the abbreviation of the aggregation. The abbreviations of aggregations are denoted in parentheses of each aggregation's name, while the abbreviations of measures' names are just above. For instance, the seeding strategy named *InExp* refers to the structural measure in-degree and aggregation exponential forgetting presented in Equation 6.14, while *BetPowMin* is the combination of betweenness and minimum power introduced in Equation 6.10.

6.7 Experimental Environment

6.7.1 Introduction

The goal of the experimental study is to determine whether the proposed *tInf* algorithm for maximizing the spread of influence SI in temporal social networks is able to provide satisfying results within an acceptable time frame. Compared to the greedy algorithm presented in Section 4.4.3, this algorithm does not perform any kind of greedy search. It rather tries to benefit from both: potential optimal value of K and time-sensitive network measures which were then the base for ordering nodes and choosing seeds.

In order to provide a detailed experimental study, a general strategy is applied. Instead of analysing multiple factors influencing the results concurrently, they are analysed separately to show to what extent they are important in the process of spread of influence. Moreover, they serve as justification for the values of parameters selected for the whole experimental study, such as budget c, threshold θ or seeding strategy. Firstly, the role of seed-set size is analysed (see Section 6.8.1. This experiment aims to show the importance of seeding strategies if the budget c is variable. For instance, it may be not needed to look for time-consuming seeding strategies if a less computationally complex strategy will provide the same results for a given budget c. Next, in Section 6.8.2 three different values of the threshold θ are analysed to see how the overall results look and how they differ when θ is changing. Next, the strategies of choosing seeds are evaluated in Section 6.8.3. In

order to not to overload the main experimental part of this thesis, only the two best among them were selected for the next and most important part of the study which is the evaluation of the algorithm *tInf*. It is presented in Section 6.8.4 and here this method is evaluated for fixed budget c, fixed threshold parameter θ, all datasets D_z and all the shifts u, two best seeding strategies, as well as for all the temporal networks generated in Section 5.2.4. Moreover, it is compared against two random strategies and the greedy algorithm. Lastly, in Section 6.8.5, the computational complexity and time of computations is compared.

In the following subsections all the necessary details of experimental study are described. Some of them were introduced and extensively discussed in Section 5.2, but in order to provide the complete set of information, those are also referred to here.

Additionally, the summary of how the whole process looks is presented in Figure 4.2.

6.7.2 Social Influence Model

The model discussed in this thesis is the Linear Threshold model (see Section 3.4.3.2). This model requires two factors defined: the threshold θ and weights w. Threshold θ represents the weighted threshold of neighbours of the node v_i that have to be influenced in order to make this node influenced. In this experimental scenario, θ is the same for all nodes: 0.75. The second factor is the weights over edges w which represent the level of influence of a node to another node. Weights are constructed as defined in Definition 2, i.e. this is the fraction of events initiated by a node v_i to the particular recipient v_j and the number of all events initiated by v_i. It is then assumed that the intensiveness of communication represents the power of influence here. The advantage of this definition of weights is that they change as the network changes, i.e. they change from time window to time window.

6.7.3 Budget

The task of maximizing the spread of influence also requires providing the budget c which is used for influencing initial nodes, in the experimental part it is defined

as the percentage of nodes from the training period (rounded up). Those nodes form the seed set $\Phi(0)$.

6.7.4 Datasets

The experiments were conducted by using five different datasets obtained from the repository KONECT. Those datasets are described in Section 5.2.2.

6.7.5 Temporal Social Networks

As the challenge is to maximize the spread of influence in temporal social networks, they have to be introduced and defined. The definition of the temporal social network as used in this thesis is presented in Definition 4 and the configurations of them as used in the experimental study are introduced in Section 5.2.4.

For each dataset five periods of each dataset are used - each of them covers 60% of time of the dataset and they are moved for 5% from the beginning in order to introduce different starting points for the algorithms - these are the training periods. Then, for each training period the consecutive 20% of the dataset period is used as a testing period for the evaluation. For the training period a number of temporal social networks are generated, each of them consisting of K time windows, see Table 5.9. The training period is split into ten time windows and this is the number of steps in which particular algorithms will be evaluated. It is worth underlining, that the total number of influenced nodes after the process ends may be slightly higher than the number of nodes in the testing period, since the number of seeds is also taken into account, even if they do not appear later in $\{T_1, \cdots, T_{10}\}$. They have to be included in the influence set, since at the beginning of the process it is not known whether they appear in next periods or not. Hence, the upper limit for the process outcome is the number of nodes in the testing period plus the number of seeds $\Phi(0)$. This process is illustrated in Figure 4.2.

6.7.6 Seeding Strategies

The strategies for choosing seeds are defined in Section 6.6.5. Additionally, the greedy algorithm is also evaluated and two random strategies are introduced. These three methods are briefly described below.

Greedy algorithm (*Greedy*)

The greedy algorithm is relatively straightforward, see Section 4.4.3. It is developed for static networks, which is why in this experimental study it is used for a time-aggregated network ($K = 1$). This is because it is not suited to deal with temporal social networks and yet to align the *tInf* algorithm to some commonly used methods, this simplification was used. Unfortunately, due to its nature, the greedy algorithm as a computationally complex one, could run only on networks built over datasets D_1 and D_3. Yet there exist improvements to the greedy algorithm that allow to run it shorter, but at the expense of quality of results as presented in Section 4.4.4. However, it was decided to run the greedy algorithm as it was originally defined, see Algorithm 3.

Random (*Rand*)

This method draws the requested number of nodes from the testing period, without repetitions. It does not take any advantage of the temporal aspects of the generated temporal social network. In order to make the results for the random method representative, a hundred of iterations for different random seed sets are made and the results are averaged.

Random frequent (*RandFreq*)

This method selects nodes based on their frequency of occurrence in particular time windows, i.e. the node that occurs most frequently in all time windows before the seed selection will have a greater chance to be selected as a seed. Similarly to the *Rand* method, the results are averaged across a hundred iterations.

The names of those methods introduced in parentheses will be used in the experimental study to refer to these seeding strategies.

6.7.7 Hardware and Software Framework

The hardware and software framework used for the experiments are described in Section 5.2.5.1.

6.7.8 Time of Computations

Since the problem of finding a seed set which maximizes the spread of influence is NP-hard, many different strategies were proposed, see Section 4.4. In order to be able to compare the proposed algorithm with others, apart from the quality measured as a number of influenced nodes at the end of the process, also the time of computations is measured. Since the proposed method *tInf* differs from the approach used by the greedy algorithm, it was decided that the overall running time consists of two operations: building the TSN (even if it is a time-aggregated static network) and finding seeds and for *tInf* algorithm it is actually the time of execution the algorithm (see Algorithm 4). The time for the spread of influence is not measured since it is outside the responsibility of the seeding strategies.

6.8 Experimental Results

6.8.1 The Analysis of Seed-set Size

An exemplary application of the maximizing the spread of influence process is the marketing campaign in a social network. Those campaigns always have some restrictions surrounding the budget, understood as some limitation on how many nodes can be initially influenced. This budget in terms of direct marketing is spent on sending gifts, directed advertisements or on any other form of activity that can influence an individual. As it is presented in Section 4.2, in the research task of maximizing the spread of influence this budget is most often expressed as a number of nodes that will be influenced in the first period, just before the influence process starts. They form the seed set $\Phi(0)$ and most often it is assumed that the cost of influencing any person in the social network is the same for all nodes. Regarding the budget limitation, it is expected that the outcome from the process of the spread of influence will be maximised. The best initial seed set will

be chosen in order to influence as many nodes as possible. Since the social network cannot be controlled, this is the moment when particular seeding strategies may outperform others within a given budget. However, it is interesting to observe how particular strategies' results differ when the budget is changing, i.e. evaluate them in terms of varying budget c. This may provide the answer to the question of whether is it preferred to focus on computationally complex strategies that do not provide significantly better results if the budget is minimally increased. In this section there are five seeding strategies for the two temporal social networks evaluated, namely $TSN^{16}_{D_1,T_3^{Tr}}$ and $TSN^{32}_{D_3,T_4^{Tr}}$. These networks are selected, since these are the only networks that allow the seed set using the greedy algorithm to be found - other networks are too big. In the experiments budget c is used as defined in Section 6.7.3, i.e. it is the fraction of nodes from the training period rounded up. Budget c is changed from 1% to 20% and the following values are evaluated: 0.01, 0.05, 0.10, 0.15, 0.20. The seeding strategies are the following: *OutExp*, *InHyp*, *Greedy*, *Rand*, *RandFreq*. For all experiments the threshold θ is fixed for all nodes and equals 0.75.

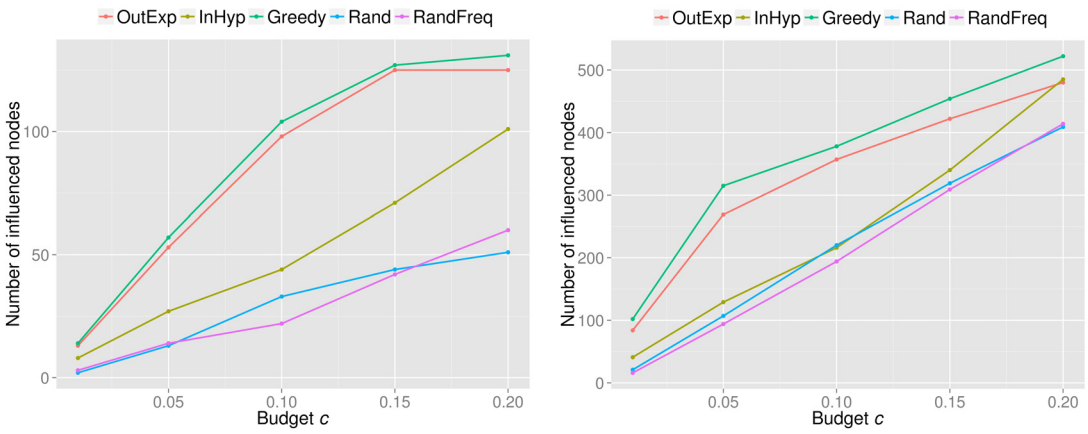
This experimental scenario is presented in Table 6.3 and the results are presented in Figure 6.7. When looking at the results, it is observed that the influence of the budget on the results is similar for $TSN^{16}_{D_1,T_3^{Tr}}$, as presented in Figure 6.7a, and $TSN^{32}_{D_3,T_4^{Tr}}$, as shown in Figure 6.7b. However, for the manufacturing company dataset the best seeding strategies, *OutExp* and *Greedy* at some point tend to have smaller sensitivity to the value of the budget. This is rather due to the fact that they are close to the upper limit of nodes in the testing period. However, the increase of the budged results with a linear increase in the number of influenced nodes and does not make seeding strategies' outcomes particularly similar, at least to $c = 0.20$. Higher values of budget c are not considered here, since it is very unlikely that over 20% of nodes will become seeds $\Phi(0)$ for typical scenarios such as running marketing campaigns.

6.8.2 The Analysis of Threshold θ

The next analysis presents how the value of the threshold determines the results. The threshold θ represents the weighted fractions of neighbours of a node that

Table 6.3: Parameters for the spread of influence process - varying budget c.

Experiment 1		
Parameter	Value	
Propagation model	Linear Threshold	
Threshold θ	0.75	
Training TSN	$TSN^{16}_{D_1,T_3^{Tr}}$	$TSN^{32}_{D_3,T_4^{Tr}}$
Testing TSN	$TSN^{10}_{D_1,T_3^{Te}}$	$TSN^{10}_{D_3,T_4^{Te}}$
Source dataset	D_1	D_3
Shift u	3	4
No. of windows - training	16	32
No. of windows - testing	10	
No. of nodes - training	157	1,542
No. of nodes - testing	136	368
Budget c	{0.01, 0.05, 0.10, 0.15, 0.20}	
Seeding strategies	{*OutExp*, *InHyp*, *Greedy*, *Rand*, *RandFreq*}	



(a) *Manufacturing company*, $D_z = 1$ (b) *University of California*, $D_z = 3$

Figure 6.7: The influence of budget c on the process outcome.

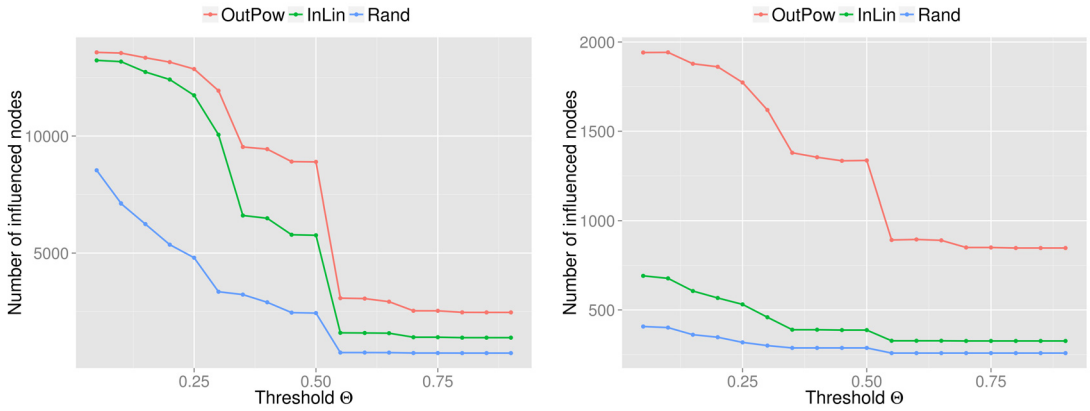
Table 6.4: Parameters for the spread of influence process - varying threshold θ.

Experiment 2		
Parameter	Value	
Propagation model	Linear Threshold	
Threshold θ	from 0.05 to 0.90 (every 0.05)	
Training TSN	$TSN^{128}_{D_4,T_3^{Tr}}$	$TSN^{64}_{D_5,T_2^{Tr}}$
Testing TSN	$TSN^{10}_{D_4,T_3^{Te}}$	$TSN^{10}_{D_5,T_2^{Te}}$
Source dataset	D_4	D_5
Shift u	3	2
No. of windows - training	128	64
No. of windows - testing	10	
No. of nodes - training	17,150	22,638
No. of nodes - testing	28,246	11,162
Budget c	0.01	
Seeding strategies	{*OutPow*, *InLin*, *Rand*}	

have to be influenced in order to activate this node (see Section 3.4.3.2 for details). Naturally, the higher the threshold, the harder it is to influence an individual, but maybe the value of the threshold does not influence the results strongly? As it is presented in Section 3.4.3.2 there are many methods for setting or estimating this value for individuals in simulation scenarios, but here, the threshold is fixed for all the nodes. The argument is that this thesis focuses on the temporal aspects of social networks rather than on the individuals' perspective. However, differentiating the threshold value for nodes in the network is planned for future works.

In order to study the influence of the threshold value on the results, another two temporal social networks are evaluated, $TSN^{128}_{D_4,T_3^{Tr}}$ and $TSN^{64}_{D_5,T_2^{Tr}}$, i.e. Facebook and Digg networks, respectively. The value of threshold θ varies from 0.05 to 0.90 every 0.05. This time the budget is fixed and $c = 0.01$. Only three seeding strategies are evaluated here, namely *OutPow*, *InLin* and *Rand*, since the greedy algorithm cannot run in reasonable time for those datasets (see Table 6.1 for argumentation). Results are presented in Figure 6.8.

It is observed that the threshold θ may significantly influence the results for the spread of influence. For the best-performing seeding strategy *Outpow* in two TSNs the threshold of 0.05 results with 5.52 and 2.29 times more influenced



(a) *Facebook*, $D_z = 4$ (b) *Digg*, $D_z = 5$

Figure 6.8: The influence of threshold θ on the process outcome.

nodes than for the threshold 0.90. Yet, for both temporal social networks, at some point the increase of the threshold does not decrease the number of influenced nodes, in both cases this is $\theta = 0.50$. This might be the absolute border of the threshold value where technically the process cannot run further due to the social network configuration, i.e. the placement of nodes and edges. Most likely, there were many nodes with just two neighbours and the same number of events aggregated into one window. In this case with a further increase of the threshold level, the weighted fraction of neighbours is too small to overcome the θ border. Those moments are the most challenging, since they also verify the quality of seeding strategies. This is the reason why in further experiments the threshold level $\theta = 0.75$ - it allows us to differentiate the seeding strategies the most.

6.8.3 The Analysis of Seeding Strategies

As previous experiments presented in Sections 6.8.1 and 6.8.2 show, there are visible differences in seeding strategies. As for now, it is seen that the greedy algorithm outperforms other evaluated methods. Definitely the worst performing are Rand and *Randfreq*, which is not especially surprising. However, at least for conducted Experiments 1 and 2, *OutExp*, *OutPow*, *InHyp* and *InLin* present at least comparable results. The second best performing measure yet, is *OutExp* which confirms the results presented in Michalski et al. [2014]. However, in that

work the *greedy* algorithm was not evaluated and, as Figures 6.7 and 6.8 present, the difference between *OutExp* and *Greedy* seems really small, at least for these four TSNs. In fact, the greedy algorithm is the one that performs well for static networks and it was not evaluated in the temporal environment until this work. It must be remembered that in this thesis the greedy is run on a time-aggregated network, since no temporal modification of that exists yet. So, this might be the true reason why it is the best, but the differences are not significant - most of the other proposed seeding strategies take advantage of the activity of nodes in time, as presented in Section 6.7.6. Moreover, knowing that the difference in run time is significant, as presented in Section 6.8.5, the advantage of the greedy algorithm becomes even weaker.

Despite that the goal of this thesis is not to evaluate seeding strategies, but to analyse the influence of temporal network configuration on the spread of influence, in this section the comparison of best-performing and random seeding strategies is presented. The experiments were conducted for all the seeding strategies and temporal social networks, but as it is argued above, as this is not the purpose of this thesis, just a selection of results is presented. And for further experiments only the top two of those best-performing seeding strategies are chosen with two random methods as a baseline.

The results of evaluating the influence of seeding strategies are presented in Table 6.5. It presents the total number of influenced nodes for a given temporal social network scenario. Those values were than compared to the total number of nodes in the testing period and presented in Figure 6.9. As it is observed, the best performing seeding strategies are *Greedy*, *OutExp* and *BetHyp*, so these will be included in the analysis of the *tInf* method presented in Section 6.8.4.

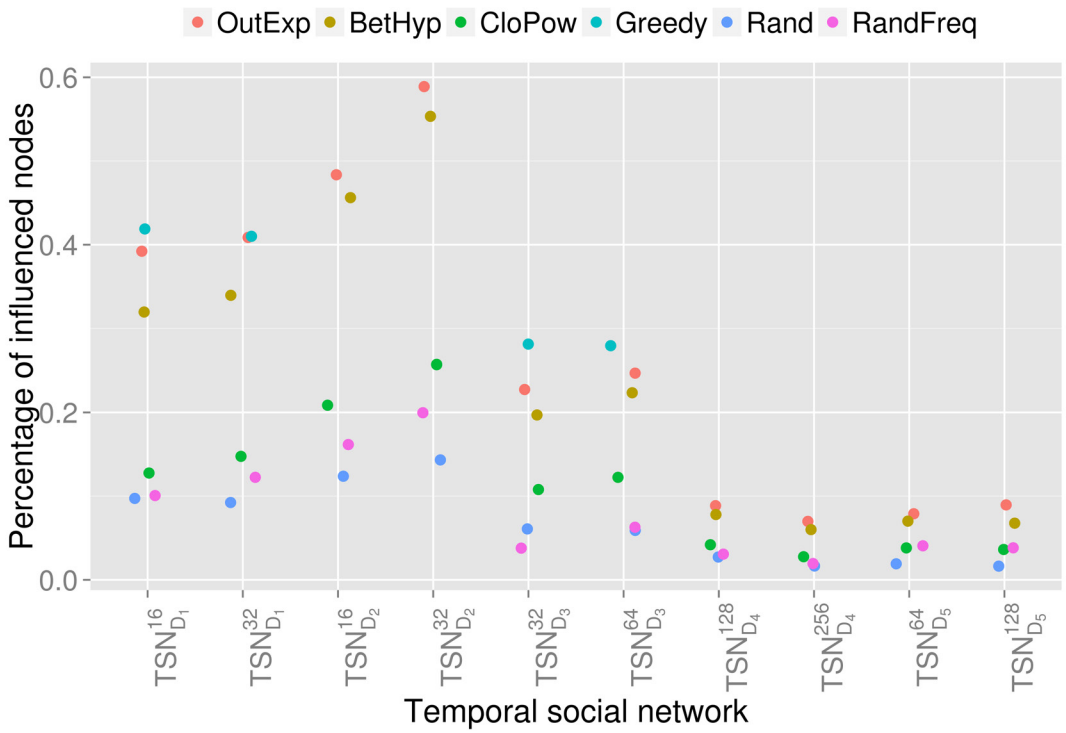


Figure 6.9: The comparison of seeding strategies for selected TSNs.

Table 6.5: The influence of seeding strategies on the process outcome.

Experiment 3										
TSN	Dataset D_z	Shift u	K	Budget c	*OutExp*	*BetHyp*	*CloPow*	*Greedy*	*Rand*	*RandFreq*
$TSN^{16}_{D_1,T^{Tr}_3}$	D_1	3	16	0.05	53	44	18	57	13	14
$TSN^{32}_{D_1,T^{Tr}_2}$	D_1	2	32	0.05	56	47	21	57	13	16
$TSN^{16}_{D_2,T^{Tr}_3}$	D_2	3	16	0.01	24731	23712	11090	-	6122	8121
$TSN^{32}_{D_2,T^{Tr}_4}$	D_2	4	32	0.01	26515	24876	11892	-	6122	9122
$TSN^{32}_{D_3,T^{Tr}_4}$	D_3	4	32	0.01	106	104	41	102	21	16
$TSN^{64}_{D_3,T^{Tr}_5}$	D_3	5	64	0.01	92	81	44	102	21	21
$TSN^{128}_{D_4,T^{Tr}_3}$	D_4	3	128	0.01	2527	2357	1219	-	719	975
$TSN^{256}_{D_4,T^{Tr}_5}$	D_4	5	256	0.01	2875	2511	1381	-	719	1012
$TSN^{64}_{D_5,T^{Tr}_2}$	D_5	2	64	0.01	1053	895	451	-	258	516
$TSN^{128}_{D_5,T^{Tr}_3}$	D_5	3	128	0.01	850	756	541	-	258	425

6.8.4 Maximizing the Spread of Influence

Finally, after conducting the experiments showing the relationship between the budget c, threshold θ, seed selection strategies and the number of influenced nodes, the proposed *tInf* method can be evaluated. Since it would be impossible to evaluate all the possible configurations, the following parameters are chosen based on the outcomes of Sections 6.8.1–6.8.3:

- Datasets: $\{D_1, \ldots, D_5\}$
- Budget $c = 0.05$ for D_1 and $c = 0.01$ for other datasets
- Threshold $\theta = 0.75$
- Seed selection strategies: *OutExp*, *BetHyp*, *Greedy*, *Rand*, *RandFreq*

The larger budget c was necessary for D_1, since if it would be 0.01, just two nodes will be seeds and the final results would not differ significantly.

The goal of this experiment is to find out how the number of influenced nodes changes according to varying K, i.e. the number of component social networks in the temporal social network. The idea of the *tInf* algorithm is to exploit this dependency and to find appropriate K to run the seed selection method in order to maximize the number of influenced nodes. Initially, the evaluated TSNs are the same as presented in Table 5.9, i.e. 270 temporal social networks are evaluated: 54 TSNs configurations with 5 shifts u each. Similarly to previous experiments, the testing period consists of ten time windows, i.e. $K' = 10$. The results are averaged across all shifts u and presented according to the averaged number of nodes in the testing period.

Table 6.6: The fraction of influenced nodes compared to the number of nodes in the testing period. If there is no value, it is because either the greedy method run too long or a particular TSN was not evaluated.

Dataset	Seeding strategy	K											
		1	2	4	8	16	32	64	128	256	512	1024	2048
D_1	*OutExp*	0.2	0.3	0.34	0.37	0.38	0.41	0.41	0.43	0.45	0.46	-	-
	BetHyp	0.19	0.22	0.25	0.28	0.32	0.34	0.38	0.41	0.44	0.45	-	-
	Greedy	0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.41	0.41	-	-
	Rand	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	0.09	-	-
	RandFreq	0.07	0.08	0.08	0.09	0.1	0.12	0.12	0.13	0.16	0.17	-	-
D_2	*OutExp*	0.46	0.46	0.47	0.49	0.53	0.56	0.57	0.59	0.61	0.6	0.58	0.56
	BetHyp	0.33	0.4	0.44	0.47	0.5	0.53	0.55	0.57	0.58	0.57	0.55	0.52
	Greedy	-	-	-	-	-	-	-	-	-	-	-	-
	Rand	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13	0.13
	RandFreq	0.16	0.17	0.17	0.17	0.17	0.19	0.2	0.2	0.2	0.2	0.2	0.16
D_3	*OutExp*	0.16	0.17	0.19	0.21	0.23	0.26	0.22	0.2	0.17	0.16	-	-
	BetHyp	0.12	0.15	0.18	0.19	0.21	0.25	0.19	0.19	0.17	0.16	-	-
	Greedy	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	-	-
	Rand	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	-	-
	RandFreq	0.04	0.04	0.04	0.04	0.05	0.06	0.05	0.04	0.04	0.04	-	-
D_4	*OutExp*	0.06	0.06	0.07	0.07	0.07	0.08	0.08	0.08	0.09	0.1	0.1	0.09
	BetHyp	0.05	0.06	0.07	0.07	0.07	0.07	0.08	0.08	0.08	0.09	0.09	0.08
	Greedy	-	-	-	-	-	-	-	-	-	-	-	-
	Rand	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
	RandFreq	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
D_5	*OutExp*	0.07	0.07	0.08	0.08	0.08	0.09	0.09	0.07	0.07	0.06	-	-
	BetHyp	0.06	0.07	0.07	0.07	0.07	0.08	0.07	0.06	0.06	0.06	-	-
	Greedy	-	-	-	-	-	-	-	-	-	-	-	-
	Rand	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	-	-
	RandFreq	0.02	0.03	0.03	0.04	0.04	0.05	0.04	0.03	0.03	0.03	-	-

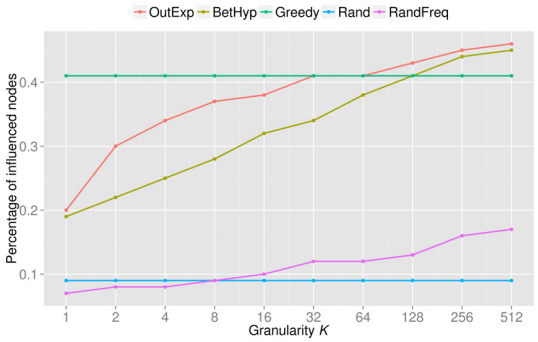
The results reveal that for the most of the time-sensitive measures the number of influenced nodes rises with the higher granularity of the underlying training temporal social network. Yet, at some point, at least for evaluated measures and datasets, the number of influenced nodes starts to reduce. This phenomenon is presented in the Figure 6.10b-e. For instance, for the Enron dataset, the peak in the number of influenced is observed for $K = 256$ and for Digg it is $K = 32$, see Figure 6.10b and 6.10e, respectively. Moreover, for the datasets, for which it was possible to evaluate the greedy algorithm, namely D_1 and D_3, while approaching the best K value, the best time-sensitive seeding strategies were able to outperform the greedy method.

Then, as described in Section 6.6.3, linear regression is run in order to find the parameters $a_1, \ldots, a_j$. The tool used to perform the regression was the KNIME software (Berthold et al. [2008]). By using this tool it is observed which input factors are the most important when choosing the most appropriate K from K^{set}. The linear regression is run separately for every dataset. The TSNs generated for four shifts serve as the input, and the TSNs generated for the fifth shift is used for evaluation, as presented in Sections 6.6.3 and 5.2.4. For those networks, a number of measures are computed and averaged, see Section 6.6.3. Table 6.7 presents the regression results for all datasets. It contains the averaged parameters for input values and in the bottom part of it, it is presented the root-mean-square error when using cross-validation (see Section 6.6.3 for details). This error demonstrates how well the regression performed for the best K found for the evaluated TSNs comparing to the real number of influenced nodes. This error is presented only for the granularity of the TSN that has been actually used for seeding among other evaluated granularities.

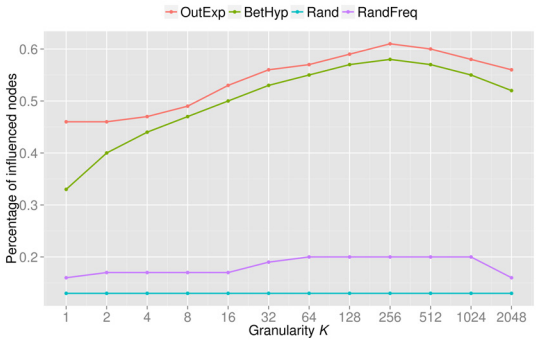
6.8.5 Run Time and Computational Complexity

6.8.5.1 Run Time

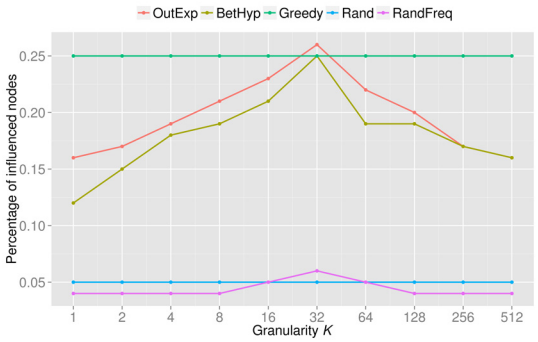
The evaluated seeding strategies may be assigned to one of two groups: using temporal information and disregarding it. The first group contains all the methods that take advantage of temporal social networks other than the time-aggregated network $K = 1$, which, in fact, represents the static network. Those seeding



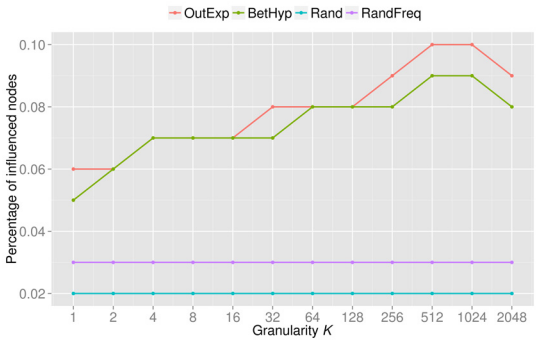
(a) *Manufacturing company*, $z = 1$



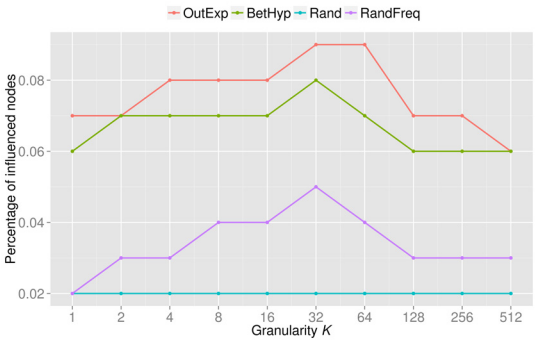
(b) *Enron*, $z = 2$



(c) *University of California*, $z = 3$



(d) *Facebook*, $z = 4$



(e) *Digg*, $z = 5$

Figure 6.10: Spread of influence for different values of K.

Table 6.7: Linear regression results for all the evaluated datasets.

	Source dataset				
Input value	D_1	D_2	D_3	D_4	D_5
AvgBetweenness	-0.0335	0.1092	0.032	-0.1231	-0.0774
AvgTotalDegree	1.3561	1.4799	1.5128	1.2839	1.4611
AvgCloseness	-0.0892	-0.023	0.21	0.1723	0.0916
AvgClustering	0.022	0.011	0.1299	0.2713	0.3369
AvgDensity	0.1651	0.091	-0.1011	-0.083	-0.1609
AvgNumberOfConnectedCompo	-0.7631	-0.7923	-0.8119	-1.001	-1.1246
Intercept	0.3475	0.004	0.311	0.2091	0.1294
Root-mean-square error	0.0188	0.053	0.0522	0.0321	0.0612

strategies require to generate a series of TSNs in order to fully benefit from the time-sensitive properties of communication. The methods in this group are all the methods enumerated in Section 6.6.5 and *GreedyFreq* introduced in Section 6.7.6. The second group for choosing seeds uses the time-aggregated network, hence the seeding strategies in this group give the same results, regardless the value of K. The methods include *Greedy* and *Rand*, see Section 6.7.6. Moreover, the *tInf* algorithm is strongly dependent on temporal social networks. It looks for a specific granularity, namely a given K, where there is a chance to maximize the spread of influence, see Algorithm 4. So, in order to use this method, a consecutive series of TSNs for different K has to be generated and, next, the selected structural measures have to be computed.

Commonly, the quality of the seeding strategies always should be accompanied by the time of computations, the same approach is used in this thesis. Ignoring the run time issues may significantly change the general impression how good those methods are. Moreover, as the seeding strategies are applied to temporal social networks, the time for choosing seeds may also be limited, since the network can change at the same time when computations are performed. As a result, the final seed set will no longer be adequate. The total running time of *tInf* consists of two factors: (i) generating the TSN and (ii) running the seeding strategy, as it is presented in Section 6.7.8. The first task, apart from creating TSNs, has to find the best K, while the second applies a given seeding strategy. This is why the computation of the the averaged measures is performed for each time window

of the TSN and the chosen seeding strategy is applied just once, at the end of this process, as Algorithm 4 denotes.

Table 6.8 presents how long it takes to generate particular TSNs used in the *tInf* algorithm. It shows whether the time depends on the value of K, the number of vertices or number of nodes. This table does not present the total time of running *tInf* algorithm or the greedy method, but aims to show the dependency of event sequence characteristics and time needed to generate the TSNs. The shift used for generating TSNs is 0%, i.e. $u = 1$. It is observed that the time needed to generate all TSNs is increasing up to 6 times when increasing the value of K. Moreover, it is more dependent on the number of events in the source TES rather than on the number of individuals. For instance, when comparing D_3 and D_5, where the number of events is similar, but the number of individuals differs, results reveal that the time needed to generate the TSNs is similar.

Table 6.8: The time needed for generating training TSNs (seconds).

Dataset index z	Number of indiv.	Number of events	Time (seconds)											
			K											
			1	2	4	8	16	32	64	128	256	512	1024	2048
1	166	50,391	17	17	20	24	30	36	47	61	80	104	-	-
2	30,331	346,339	588	484	497	530	579	654	765	908	1064	1243	1252	1293
3	1,786	54,870	5	5	5	6	7	10	14	19	26	40	-	-
4	12,045	149,227	42	41	42	43	46	50	54	62	78	95	124	173
5	21,132	47,768	7	5	4	6	6	7	9	13	19	24	-	-

The Table 6.9 presents the total running time of the *tInf* algorithm for the best seeding strategies. The seeding strategies are *OutExp* and *BetHyp*. Those are compared with the total running time of the *Greedy* algorithm. Please note that the greedy method uses a time-aggregated network, so it also has to generate a TSN with $K = 1$, but it does need to generate other TSNs. The total running time is for the scenarios presented in Table 5.9, i.e. the set K^{set} consists of the values of K as in that table. The shift u is always 1 and no optimizations for the greedy algorithm were applied in order to compare the outcome of the proposed method to the best results provided by the greedy algorithm. Since the computations were made in a distributed environment, as presented in Section 5.2.5.1, the running time is a sum of execution time for each thread. The time for *tInf* method includes the whole procedure, i.e. creating all the TSNs, narrowing down the K from a given K^{set} and computing necessary measures for all nodes in order to create the seed set $\Phi(0)$.

Table 6.9: Run time for the best evaluated seeding strategies (seconds).

Dataset index z	Number of indiv.	Number of events	The best K	Total processing time (s)		
				tInf		*Greedy*
				OutExp	*BetHyp*	
1	165	49,414	512	496.05	495.97	1,278.37
2	36,130	459,816	256	10,105.01	10,104.99	-
3	1,759	54,748	32	193.65	193.4	23.29 (days)
4	14,323	220,144	1024	1,139.26	1,139.52	-
5	22,638	53,510	32	158.86	158.86	-

It is observed that the time for the whole procedure is reasonably short for the *tInf* method, varying from 1.5 minutes (D_3) to about 3 hours (D_2). Comparing to the greedy method, *tInf* is shorter from 2.5 times (dataset D_1) up to 10,000 times (dataset D_3). Unfortunately, for other datasets it was impossible to obtain the results of the generic greedy method in reasonable time, since the number of nodes was too large in order to find the seed set effectively.

It is observed that the run time does not heavily depend on the calculating measures part of the process, because if comparing the time needed for generating TSNs as presented in Table 6.8, the majority of *tInf* running time is consumed by this part (about 80% of the total time).

6.8.5.2 Computational Complexity

To provide information about the computational complexity, it is necessary to present what are the main steps in the proposed *tInf* method:

1. Generating TSNs from TESes for all K from a given K^{set}.
2. Calculating the global network measures needed to narrow down the sub-optimal K.
3. Running the seeding strategy to select $\Phi(0)$.

The total time of *tInf* algorithm consists of these three components. For the first one, the computational complexity depends on the number of individuals and events, as well as on the K^{Σ}, where K^{Σ} is the sum of all $K \in K^{set}$. The computational complexity of this step is $\mathcal{O}(|K^{\Sigma}| * |V| * |EV|)$, where $|V|$ is the number of nodes and $|EV|$ denotes the number of events.

When computing the structural measures in order to find the value of K resulting with potentially the best spread of influence SI, the complexity of computing the individual measures has to be respected. As a result, only the most computationally complex calculations need to be considered. Finally, the complexity of this step is $\mathcal{O}(|V| * |E| + |V|^2 \log |V|)$. The former element of the sum is related to the computing of density and clustering coefficient, while the latter applies to betweenness (for the fastest known algorithm, Brandes [2001]). The computational complexity of calculating the average degree and the number of connected components is negligible. Since these operations are being run K^{Σ} times, the joint complexity is as follows: $\mathcal{O}(|K^{\Sigma}| * (|V| * |EV| + |V| * |E| + |V|^2 \log |V|))$.

The last, third part, is less important in terms of computational complexity, since for the purpose of seed selection the measures were already calculated in the second step.

Concluding, the overall computational complexity of the proposed method is as follows: $\mathcal{O}(|K^{\Sigma}| * (|V| * |EV| + |V| * |E| + |V|^2 \log |V|))$.

6.9 Discussion

This experimental study shows that by introducing time-sensitive seeding strategies altogether with proper TSN granularity, it is possible to outperform the costly *Greedy* strategy. Naturally, the most important question is how to find this *proper* granularity, since this is the crucial aspect of shortening the computation time. In the proposed *tInf* method, this task has been accomplished by using the linear regression model (see Section 6.6.3) or the classification model (see Section 6.6.4). It has been shown that the *tInf* approach gives acceptable results in terms of the spread of influence, while it significantly reduces the time needed for seeding.

The experimental research was very extensive and included nearly 32,000 runs of the spread of influence process using a maximum of 450 cluster computing cores in parallel (20 cluster nodes). Such set of experiments allowed to identify the most important factors that affect the results of the spread of influence. These results are more in-depth discussed below.

Firstly, in Section 6.8.1 it is shown that when increasing the seed-set size linearly, the spread of influence also increases in the same way. This conclusion is rather pessimistic, since it means that there are no spectacular gains in the SI when spending more on the budget used for seeding. Ideally, some phase transition would be expected when after reaching some value of the budget, the spread of influence will rapidly rise. This behaviour is observed in static multiplex networks, as presented in Michalski et al. [2013]. Yet, experiments conducted in this dissertation show that this phenomenon is not valid in the analysed temporal social networks.

Next, in Section 6.8.2 it is shown that the threshold θ in the LT model is important and strongly influences the results. Moreover, this threshold is not correlated linearly with the spread of influence, but it has got some crucial points, e.g. $\theta = 0.5$. It is mostly because when splitting the network more and more (i.e. while increasing K), there are many nodes with just two neighbours and the same number of events, i.e. with the same influence weight $b_{v,w}$, but only one of them is influenced (see Section 3.4.3.2 for details). The decrease of the threshold below 0.5 instantly turns all such nodes into influenced state by their activated neighbours. If $\theta = 0.5$ than an important barrier in the spread of influence in the

evaluated scenarios is not exceeded. However, θ setting cannot be controlled in the real world, since it is an internal factor of an individual. It may vary from one to another.

In Section 6.8.3, it is presented how different seeding strategies proposed in this dissertation differ in terms of the number of influenced nodes. The results clearly show that there are leading approaches that use the time-respecting methods combined with betweenness centrality or out-degree centrality as component structural measures. It means that the seeds selected based on their highest value of the structural measure usually provide the best spread of influence. It is also shown that the *forgetting* methods that respect the recent events more than the old ones in a way that they favour the most recent events more than linearly, provide significantly better results. It means that the nodes that recently gained a higher status in terms of structural measure are definitely better than the ones that used to be important in the past.

Finally, Section 6.8.4 evaluates the proposed *tInf* method for maximizing the spread of influence. It has been evaluated for 270 temporal social networks and for all the seeding strategies introduced in Section 6.6.5 as well as for component measures introduced in Section 5.3.3.2. It is shown that for all the temporal social networks generated, as presented in Section 5.2.4, there is a particular granularity K that outperforms the seeding applied for other K. Moreover, as it is presented in Table 6.6 and in Figure 6.10, the same granularity K^{best} is the best for all evaluated time-respecting seeding strategies. That means that this phenomenon is independent to some extent from the seeding strategy applied and it is more related to the temporal granularity itself rather than to a particular seeding strategy.

The results reveal that there are significant differences in the final fraction of influenced nodes for different datasets and it varies from 6% for D_5 to 61% for D_2, as presented in Table 6.6. This is because of different properties of the evaluated datasets. The dataset D_5 consists of the largest number of connected components, so it was impossible to influence any nodes in the components that did not contain any seed nodes.

Since the threshold $\theta = 0.75$, it was hard to overcome this exorbitant limit for some datasets, such as for D_4 or D_5. This resulted in a low fraction of influ-

enced nodes, even though the new seeding strategies still performed better than the random approaches. For the datasets for which it was possible to run the *greedy* method, the results revealed that the best time-respecting seeding strategies always outperformed the greedy method for the best K. Naturally, the *greedy* method was originally suited to static networks, and the new methods operate in the dynamic environment. However, both environments are not equivalent and it is impossible to compare both methods authoritatively. Yet, this analysis shows that instead of using computationally complex methods, some simpler solutions may perform well. Naturally, as it was presented in Sections 4.4 and 5.2.1, the temporal networks still suffer from a lack of commonly agreed models for the temporal spread of influence.

The results of the linear regression show that it is possible to find the sub-optimal values of K - the network granularity that is the best for seeding, see Table 6.7. The greatest positive influence on the SI is due to the average total degree measure (the coefficient greater than 1.2). The negative impact has the average number of connected components, the coefficient less than -0.75. The coefficients for other five global structural measures are from -0.16 to 0.35. Yet, it would be worth studying how the spread of influence behaves if more values of K could be evaluated, especially the values close to the best K identified so far. Unfortunately, the resource constraints made this task impossible.

The experimental results revealed that probably the most important aspect of temporal granularity is that for greater K the conversations may be isolated within some components in separate social networks. The time-respecting seeding strategies pick the leaders of these conversations in a chronological order. For instance, if two individuals were equally important in two conversations, the one from the recent conversation will be place higher in the ranking used for seeding.

Finally, the analysis of run time and computational complexity shows that the proposed *tInf* method does not suffer from the problem of long run time. Compared to the *greedy* method, the *tInf* run time can be several orders of magnitude shorter, see Table 6.9. The greedy algorithm runs from 2.5 for the dataset D_1 up to 10,000 times longer (D_3) than the proposed *tInf* method. Moreover, the greedy method working on the datasets with around 1800 of nodes needs about a month in the cluster environment; this is why it could not be run for the larger

datasets in a reasonable time.

Naturally, this research does not provide answers to all the questions related to the proposed method. Since just the experimental part of the study took nine months, it was hard to cover all the relevant realms. Nevertheless, compared to the evaluated methods or strategies, these questions appear to be minor, but worth enumerating. They are presented in the last chapter of this dissertation, namely Chapter 7, as future work directions. They may be summarized as follows:

- we need to know how far should we look into the past in order to maximize the spread of influence for the future
- it is expected to introduce native temporal models for the spread of influence that will substitute generic models, such as LT or IC
- the need for proper alignment of the spread of influence models' parameters to real world cases remains vital.

The proposed *tInf* method suffers from some drawbacks, even though they are not significant. It makes the *tInf* method a good progress compared to all other approaches known so far in this area. The most important drawback is that it is based on a model of the spread of influence that does not respect the time factor or the temporal information. Moreover, the introduction of the time-respecting models of the spread of influence will prevent from the simplification of the spreading process (evaluation period).

Yet, as the analysis of the state-of-the-art shows, it is expected that in the near future research in the area of the spread of influence will switch to the temporal scenarios. The reason for this is simple: they are closer to the real world settings.

6.10 Conclusions

As it is presented in this chapter, the general conclusion of the thesis is that the temporal properties of the social networks may be effectively used to maximize the spread of influence. The proposed *tInf* method takes into account the historical changes in the social network for seeding (learning case) to reach greater spread of influence SI for the future. By proposing the new *tInf* method it was possible

to observe that for some configurations, namely for different values of granularity K, the spread of influence may be even greater compared to the commonly used time-aggregating methods. In opposite to the greedy algorithm, it was achieved without involving heavy computational procedures, i.e. by taking advantage of the appropriate configuration of the temporal social network used for seeding. The time-respecting seeding strategies proposed in this chapter show that, as expected, the most influential leaders are the ones that have been important recently, not in the past. Moreover, if using reasoning method (e.g. linear regression), it was possible to calculate the final spread of influence SI for a given K very quickly. Further, for a given set of K (K^{set}), it is feasible to select the best K without running the spread of influence. Because of the speed of the proposed *tInf* method, it may be applicable even if a prompt decision about potential seeds is expected.

Chapter 7

Summary and Future Work

This thesis aims to present how to benefit from the temporal information embedded in the social network dynamics instead of ignoring it. It was presented that the problem of maximizing the spread of influence applied to the temporal networks may be successfully solved in reasonable time. Since, it is questionable to stick to the static approach for analysing dynamic processes in networks, the temporal case should be considered instead. The main goal of the new *tInf* method is to provide a scalable and efficient solution for maximizing the spread of influence SI in temporal social networks TSN, making it a reasonable choice for dynamic configurations.

The general conclusion is that the social networks around us are not static, but in fact dynamic. Since it was presented in Section 6.5, ignoring this fact is an unacceptable simplification leading to wrong conclusions about the real world. Naturally, while looking at the model introduced in this thesis in Section 6.3.2, the static network can be still used for seed selection, but as the experiments have shown, the results may be better if the temporal approach is used. Overall, the evaluation should be performed only by using the temporal network, since it more naturally reflects the real scenario. The question remains how to model the future, because there exists no commonly agreed temporal variant of the Linear Threshold model. Yet, definitely the static approach is far too simple scenario for evaluation. One of the most important arguments as to why the static greedy method can be outperformed by the time-sensitive approaches is that it does not include the time dimension at all. For the greedy method, the node that was a

very good influencer in the past may still be so in the future, ignoring the fact that the node was missing for the next part of the training period.

The experimental studies have shown that the *tInf* method can outperform the greedy algorithm. Moreover, even for non-optimal values of K the results of *tInf* are very close to the greedy method. Simultaneously, *tInf* is more efficient in terms of processing time for any K. It means that the seeding strategies may be applied efficiently to real world networks with satisfying results, even if the networks are of huge size (millions of edges).

To summarize, the contributions of this dissertation are presented below ordered by their importance - from the most to the least important.

The novel model of temporal spread of influence

This is the new way of showing how to deal with the spread of influence in temporal social networks. This novel model introduces the split of the datasets into training and testing periods, uses the temporal social network to evaluate the seeding strategies and provides a universal framework for the whole process.

The new *tInf* method for maximizing the spread of influence

The proposed *tInf* method allows one to take advantage of the temporal information about nodes' past activity in order to choose proper seed nodes to maximize the spread in the future. Its idea originates from observation of how the spread of influence changes when introducing different configurations of temporal networks. After that, it was noticed that for given criteria it is possible to increase the spread of influence and it may be obtained without huge computational effort, as presented in Chapter 6. The correlation between a spread of influence and these criteria has been experimentally supported.

Temporal structural measures

Several new temporal structural measures have been proposed and analysed in this dissertation. They can be used to evaluate node usability and node ranking for the purpose of seed selection, as presented in Section 6.6.5.

Experimental verification of the *tInf* method

The proposed method was experimentally evaluated on five real world datasets and compared against the greedy method commonly used for static networks. The results have shown the superiority of the proposed approach, at least for the cases where it was possible to compute the greedy method results.

Setting up the domain

This dissertation also aims to introduce common formal definitions for new and differently used terms, such as event sequence, temporal social network, and spread of influence in static and temporal social network, see Chapter 2. It is an attempt to provide a common understanding of the terms being used in the research in this area.

The software framework for spread of influence in temporal SNs

In order to conduct the experiments, a software framework has been developed. It allows the creation of temporal social networks of different kinds, to use the model for temporal spread of influence and to run the extensive experiments. It is based on the R language together with the *igraph* library. This framework is suitable to perform the experiments in the cluster environment which reduces the total computation time.

The results reveal that there exist leading strategies using the time-respecting methods based on betweenness or out-degree centrality. The seed set selected based on their highest value usually provides the greatest spread of influence. It has been also shown that the methods that respect the recent events much more than the old ones provide significantly better outcome.

This research revealed that a number of directions should be further examined in order to fully benefit from the proposed approach.

Firstly, there is the potential of core nodes in the network shown in Section 5.3.2.2. These are the nodes that occur in the network for the whole learning period and may be a good transmission point to spread the influence. They do not waste the budget spent on them since they are more likely to occur in the network during the following, testing periods.

Next, as the training period was constant, it is worth observing the influence of its shortening or extension on the results. Is it really needed to track such a long activity of nodes or whether it would be more beneficial to shorten the period and reduce costs?

As individuals may have different levels at which they become influenced, the next step is to relax the assumption of fixed thresholds. They may be of some distribution or randomly assigned - the idea is to verify to what extent it influences the results. One of the fascinating ideas is to change the weights of influence and thresholds in time. It has not been considered anywhere yet.

Lastly, is there any chance to develop a greedy algorithm for temporal social networks that incorporates the time factor?

As it was presented, the temporal approach provides some benefits, but also new challenges. By answering the above questions it may be possible to improve the *tInf* method to provide even better results.

Appendix A

This Appendix contains a list of published papers in which the author was considering scientific problems related to the scope of this dissertation.

Articles in JCR-listed Journals (with Impact Factor):

1. Michalski, R., Kajdanowicz, T., Bródka, P., Kazienko, P.: Seed Selection for Spread of Influence in Social Networks: Temporal vs. Static Approach. To appear in New Generation Computing in August/September 2014. Ohmsha-Japan and Springer (2014)

2. Kajdanowicz, T., Michalski, R., Musiał K., Kazienko, P.: Learning in Unlabelled Networks - An Active Learning and Inference Approach. AI Communications, in reviews since February 2014, IOS Press (2014)

Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics):

1. Jankowski, J., Kozielski, M., Filipowski, W., Michalski, R.: The Diffusion of Viral Content in Multi-layered Social Networks. ICCCI 2013, The 5th International Conference on Computational Collective Intelligence Technologies and Applications. Lecture Notes in Artificial Intelligence LNAI, vol. 8083, pp. 30-39, Springer, Berlin Heidelberg (2013)

2. Jankowski, J., Ciuberek, S., Zbieg, A., Michalski, R.: Studying Paths of Participation in Viral Diffusion Process. SocInfo 2012 - The 4th International Conference on Social Informatics. Lecture Notes in Computer Science LNCS, vol. 7710, pp. 503-516, Springer, Berlin Heidelberg (2012)

3. Jankowski, J., Michalski, R., Kazienko, P.: The Multidimensional Study of Viral Campaigns as Branching Processes. SocInfo 2012 - The 4th International Conference on Social Informatics. Lecture Notes in Computer Science LNCS, vol. 7710, pp. 462-474, Springer, Berlin Heidelberg (2012)

4. Michalski, R., Bródka, P., Palus, S., Kazienko, P., Juszczyszyn, K.: Modelling Social Network Evolution. SocInfo 2011, The Third International Conference on Social Informatics. Lecture Notes in Computer Science LNCS, vol. 6984, pp. 283-286, Springer, Berlin Heidelberg (2011)

5. Kazienko, P., Michalski, R., Palus, S.: Social Network Analysis as a Tool for Improving Enterprise Architecture. KES-AMSTA 2011, The 5th International KES Symposium on Agents and Multi-agent Systems - Technologies and Applications. Lecture Notes in Artificial Intelligence LNAI, vol. 6682, pp. 651-660, Springer, Berlin Heidelberg (2011)

Peer-reviewed Conference Proceedings:

1. Michalski, R., Kazienko P., Kajdanowicz T., Bródka P.: Data-driven Seed Selection for Spread of Influence in Temporal Social Networks. Workshop on Sociophysics at SigmaPhi 2014 - The International Conference on Statistical Physics 2014, Kaniadakis G., Scarfone A.M. (eds.), p. 75 (2014)

2. Michalski, R., Kazienko, P., Jankowski, J.: Convince a Dozen More and Succeed - The Influence in Multi-layered Social Networks. The Second Workshop on Complex Networks and their Applications at SITIS 2013 - The 9th International Conference on Signal Image Technology & Internet based Systems, December 2-5 2013, Kyoto, Japan, IEEE Computer Society, pp. 499-505 (2013)

3. Jankowski, J., Michalski, R., Kazienko, P.: Compensatory Seeding in Networks with Varying Availability of Nodes. ASONAM 2013, The 2013 IEEE and ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE Computer Society, pp. 1242-1249 (2013)

4. Kajdanowicz, T., Michalski, R., Musiał-Gabryś, K., Kazienko, P.: Active Learning and Inference Method for Within Network Classification. ASONAM 2013, The 2013 IEEE and ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE Computer Society, pp. 1299-1306

5. Kajdanowicz, T., Michalski, R., Bródka, P.: From Data to Human Behaviour. SOCIETY 2013 - International Conference on Social Intelligence and Technology, IEEE Computer Society, pp. 100-108 (2013)

6. Michalski, R., Bródka, P., Kazienko, P., Juszczyszyn, K.: Quantifying Social Network Dynamics. CASoN 2012, The Fourth International Conference on Computational Aspects of Social Networks. IEEE Computer Society, pp. 69-74 (2012)

7. Michalski, R., Jankowski, J., Kazienko, P.: Negative Effects of Incentivised Viral Campaigns for Activity in Social Networks. SCA 2012, The 2nd International Conference on Social Computing and its Applications. IEEE Computer Society, pp. 391-398 (2012)

8. Zbieg, A., Żak, B., Jankowski, J., Michalski, R., Ciuberek, S.: Studying Diffusion of Viral Content at Dyadic Level. ASONAM 2012, The 2012 IEEE and ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE Computer Society, pp. 1291-1297 (2012)

9. Michalski, R., Jankowski J., Bródka P., Kazienko P.: The Same Network - Different Communities? The Multidimensional Study of Groups in the Cyberspace. ASONAM 2014, The 2014 IEEE and ACM International Conference on Advances in Social Networks Analysis and Mining. To appear in IEEE Computer Society.

10. Michalski, R., Kazienko, P., Król, D.: Predicting Social Network Measures using Machine Learning Approach. ASONAM 2012, The 2012 IEEE and ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE Computer Society, pp. 1088-1091 (2012)

11. Michalski, R., Palus, S., Kazienko, P.: Matching Organizational Structure and Social Network Extracted from Email Communication. BIS 2011, 14th International Conference on Business Information Systems. Lecture Notes in Business Information Processing LNBIP, vol. 87, pp. 197-206, Springer, Berlin Heidelberg (2011)

Chapters in Books:

1. Michalski, R., Kazienko, P.: Maximizing Social Influence in Real-World Networks - the State of the Art and Current Challenges. To appear in Król, D., Fay, D., Gabryś, B. (eds.) Propagation Phenomena in Real World Networks, Intelligent Systems Reference Library. Springer (2014)

2. Michalski, R., Kazienko, P.: Social Network Analysis in Organizational Structures Evaluation. To appear in Encyclopedia of Social Network Analysis and Mining. Springer (2014)

3. Palus, S., Kazienko, P., Michalski, R.: Evaluation of Corporate Structure Based on Social Network Analysis. Cakir, A., De Pablos, P.O. (eds.) Social Development and High Technology Industries: Strategies and Applications, pp. 58-69. IGI-Global (2012)

References

Charu Aggarwal and Karthik Subbian. Evolutionary network analysis: A survey. *ACM Computing Surveys*, 2014. 74

Charu C Aggarwal, Shuyang Lin, and S Yu Philip. On influential node discovery in dynamic social networks. In *SDM*, pages 636–647. SIAM, 2012. 70, 71, 73

C Norman Alexander Jr. A method for processing sociometric data. *Sociometry*, 1963. 114

Andrea Apolloni, Karthik Channakeshava, Lisa Durbeck, Maleq Khan, Chris Kuhlman, Bryan Lewis, and Samarth Swarup. A study of information diffusion over a realistic social network model. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 675–682. IEEE, 2009. 23

Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012. 37, 50

Elliot Aronson. *The social animal*. Macmillan, 2003. 34, 35

Richard P Bagozzi and Kyu-Hyun Lee. Multiple routes for social influence: The role of compliance, internalization, and social identity. *Social Psychology Quarterly*, pages 226–247, 2002. 34

Norman TJ Bailey et al. *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975. 23

Albert-László Barabási. *Bursts: the hidden patterns behind everything we do, from your e-mail to bloody crusades*. Penguin, 2010. 14, 15, 56, 93

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999. 111, 113

Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. Topic-aware social influence propagation models. *Knowledge and information systems*, 37(3):555–584, 2013. 40, 41

Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11): 3747–3752, 2004. 15

Alain Barrat, Marc Barthelemy, and Alessandro Vespignani. *Dynamical processes on complex networks*, volume 574. Cambridge University Press Cambridge, 2008. 19

David J Bartholomew and David J Bartholomew. *Stochastic models for social processes*. Wiley New York, 1967. 21

Frank M Bass. A new product growth for model consumer durables. *Management Science*, 15(5):215–227, 1969. 30

Frank M Bass, Trichy V Krishnan, and Dipak C Jain. Why the bass model fits without decision variables. *Marketing science*, 13(3):203–223, 1994. 31

Alex Bavelas. Communication patterns in task-oriented groups. *Journal of the acoustical society of America*, 1950. 115

Eli Berger. Dynamic monopolies of constant size. *Journal of Combinatorial Theory, Series B*, 83(2):191–200, 2001. 39

Michael R Berthold, Nicolas Cebron, Fabian Dill, Thomas R Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, and Bernd Wiswedel. *KNIME: The Konstanz information miner*. Springer, 2008. 162

Shishir Bharathi, David Kempe, and Mahyar Salek. Competitive influence maximization in social networks. In *Internet and Network Economics*, pages 306–311. Springer, 2007. 42

Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006. 145

Francesco Bonchi. Influence propagation in social networks: A data mining perspective. *IEEE Intelligent Informatics Bulletin*, 12(1):8–16, 2011. 55, 61, 70

Allan Borodin, Yuval Filmus, and Joel Oren. Threshold models for competitive influence in social networks. In *Internet and Network Economics*, pages 539–550. Springer, 2010. 40

Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001. 168

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 145

Piotr Bródka, Stanisław Saganowski, and Przemysław Kazienko. Ged: the method for group evolution discovery in social networks. *Social Network Analysis and Mining*, 3(1):1–14, 2013. 10

Carter T Butts. Revisiting the foundations of network analysis. *science*, 325 (5939):414–416, 2009. 10

Kathleen Carley. Dynamic network analysis. *Social Networks*, 4:26, 2010. 74

Curtis R Carlson and William W Wilmot. *Innovation: The five disciplines for creating what customers want*. Random House LLC, 2006. 25

Shai Carmi, Shlomo Havlin, Scott Kirkpatrick, Yuval Shavitt, and Eran Shir. A model of internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences*, 104(27):11150–11154, 2007. 66

Peter J Carrington, John Scott, and Stanley Wasserman. *Models and methods in social network analysis*. Cambridge university press, 2005. 15, 30, 31, 32, 114

Wei Chen, Yajun Wang, and Siyu Yang. Efficient influence maximization in social networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208. ACM, 2009. 62, 64, 65, 73

Wei Chen, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1029–1038. ACM, 2010. 65, 66, 73

Wei Chen, Wei Lu, and Ning Zhang. Time-critical influence maximization in social networks with time-delayed diffusion process. *arXiv preprint arXiv:1204.3074*, 2012. 53

Robert B Cialdini. *Influence: Science and practice*, volume 4. Allyn and Bacon Boston, MA, 2001. 35

Robert B Cialdini and Noah J Goldstein. Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 55:591–621, 2004. 34

Peter Clifford and Aidan Sudbury. A model for spatial conflict. *Biometrika*, 60 (3):581–588, 1973. 42

Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5), 2006. 102

Luca DallAsta, Andrea Baronchelli, Alain Barrat, and Vittorio Loreto. Nonequilibrium dynamics of language games on complex networks. *Physical Review E*, 74(3):036105, 2006. 42

Munmun De Choudhury, Hari Sundaram, Ajita John, and Dorée Duncan Seligmann. Social synchrony: Predicting mimicry of user actions in online social media. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pages 151–158. IEEE, 2009. 89

Klaus Dietz. Epidemics and rumours: A survey. *Journal of the Royal Statistical Society. Series A (General)*, pages 505–528, 1967. 23

Peter Sheridan Dodds, Roby Muhamad, and Duncan J Watts. An experimental study of search in global social networks. *science*, 301(5634):827–829, 2003. 20

Wellesley Dodds. An application of the bass model in long-term new product forecasting. *Journal of Marketing Research (JMR)*, 10(3), 1973. 31

Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM, 2001. 61, 73

Grzegorz Dralus and Jerzy Świątek. Static and dynamic complex models: comparison and application to chemical systems. *Kybernetes*, 38(7/8):1198–1215, 2009. 71

Adam G Dunn and Blanca Gallego. Diffusion of competing innovations: The effects of network structure on the provision of healthcare. *Journal of Artificial Societies & Social Simulation*, 13(4), 2010. 33

Richard Durrett, Richard Durrett, Richard Durrett, and Richard Durrett. *Lecture notes on particle systems and percolation*. Wadsworth & Brooks/Cole Advanced Books & Software, 1988. 40

Paul Erdos and Alfréd Rényi. {On the evolution of random graphs}. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960. 113

Serge Galam. Modelling rumors: the no plane pentagon french hoax case. *Physica A: Statistical Mechanics and Its Applications*, 320:571–580, 2003. 23

Ricardo F Garcia. Birth-death processes. *Network Modeling, Simulation, and Analysis*, 61:83, 1990. 21

Jacob Goldenberg, Barak Libai, and Eitan Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing letters*, 12 (3):211–223, 2001. 40

Sergio Gomez, Albert Diaz-Guilera, Jesus Gomez-Gardeñes, Conrad J Perez-Vicente, Yamir Moreno, and Alex Arenas. Diffusion dynamics on multiplex networks. *Physical review letters*, 110(2):028701, 2013. 21

Amit Goyal, Francesco Bonchi, and Laks VS Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM, 2010. 39, 68, 69

Amit Goyal, Francesco Bonchi, and Laks VS Lakshmanan. A data-based approach to social influence maximization. *Proceedings of the VLDB Endowment*, 5(1): 73–84, 2011a. 69

Amit Goyal, Wei Lu, and Laks VS Lakshmanan. Celf++: optimizing the greedy algorithm for influence maximization in social networks. In *Proceedings of the 20th international conference companion on World wide web*, pages 47–48. ACM, 2011b. 64, 73

Amit Goyal, Wei Lu, and Laks VS Lakshmanan. Simpath: An efficient algorithm for influence maximization under the linear threshold model. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 211–220. IEEE, 2011c. 66

Amit Goyal, Francesco Bonchi, Laks VS Lakshmanan, and Suresh Venkatasubramanian. On minimizing budget and time in influence propagation over social networks. *Social Network Analysis and Mining*, 3(2):179–192, 2013. 51, 54

Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420, 1978. 38, 39, 80

Daniel Gruhl, Ramanathan Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web*, pages 491–501. ACM, 2004. 20, 23

Peter Hedström and Peter Bearman. *The Oxford handbook of analytical sociology*. Oxford University Press, 2009. 36, 39, 56

Uwe Helmke, John B Moore, and Würzburg Germany. Optimization and dynamical systems. 1994. 71

Richard A Holley and Thomas M Liggett. Ergodic theorems for weakly interacting infinite systems and the voter model. *The annals of probability*, pages 643–663, 1975. 42

Petter Holme and Jari Saramäki. Temporal networks. *Physics reports*, 519(3): 97–125, 2012. 3, 10, 15, 16, 56, 80, 113

John E Hopcroft and Robert E Tarjan. Efficient algorithms for graph manipulation. 1971. 116

José Luis Iribarren and Esteban Moro. Branching dynamics of viral information spreading. *Physical Review E*, 84(4):046116, 2011. 22

Jarosław Jankowski, Sylwia Ciuberek, Anita Zbieg, and Radosław Michalski. Studying paths of participation in viral diffusion process. In *Social Informatics*, pages 503–516. Springer Berlin Heidelberg, 2012a. 42

Jarosław Jankowski, Radosław Michalski, and Przemysław Kazienko. The multidimensional study of viral campaigns as branching processes. In *Social Informatics*, pages 462–474. Springer, 2012b. 22

Jarosław Jankowski, Michał Kozielski, Wojciech Filipowski, and Radosław Michalski. The diffusion of viral content in multi-layered social networks. In *Computational Collective Intelligence. Technologies and Applications*, pages 30–39. Springer Berlin Heidelberg, 2013a. 21

Jarosław Jankowski, Radosław Michalski, and Przemysław Kazienko. Compensatory seeding in networks with varying avaliability of nodes. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1242–1249. ACM, 2013b. 71, 73

Marco Alberto Javarone. Social influences in the voter model: the role of conformity. *arXiv preprint arXiv:1401.0839*, 2014. 42

Richard Jensen. Innovation adoption and diffusion when there are competing innovations. *Journal of Economic Theory*, 29(1):161–171, 1983. 33

Qingye Jiang, Guojie Song, Gao Cong, Yu Wang, Wenjun Si, and Kunqing Xie. Simulated annealing based influence maximization in social networks. In *AAAI*, 2011. 65

Tomasz Kajdanowicz, Radosław Michalski, Katarzyna Musiał, and Przemysław Kazienko. Active learning and inference method for within network classification. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1299–1306. ACM, 2013. 55

Tomasz Kajdanowicz, Radosław Michalski, Katarzyna Musiał, and Przemysław Kazienko. Learning in unlabelled networks - an active learning and inference approach. *AI Communications, in reviews since February 2014*, 2014. 55

Fariba Karimi and Petter Holme. Threshold model of cascades in temporal networks. *arXiv preprint arXiv:1207.1206*, 2012. 80

Przemysław Kazienko, Katarzyna Musiał, and Aleksander Zgrzywa. Evaluation of node position based on email communication. *Control & Cybernetics*, 38(1), 2009. 7

Przemysław Kazienko, Piotr Bródka, and Katarzyna Musial. Individual neighbourhood exploration in complex multi-layered social network. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 3, pages 5–8. IEEE, 2010. 21

Przemysław Kazienko, Radosław Michalski, and Sebastian Palus. Social network analysis as a tool for improving enterprise architecture. In *Agent and Multi-Agent Systems: Technologies and Applications*, pages 651–660. Springer Berlin Heidelberg, 2011. 15

Herbert C Kelman. Compliance, identification, and internalization: Three processes of attitude change. *Journal of conflict resolution*, pages 51–60, 1958. 34

Herbert C Kelman. Processes of opinion change. *Public opinion quarterly*, 25(1): 57–78, 1961. 34

David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003. 38, 40, 52, 55, 56, 61, 62, 73, 76, 128

David Kempe, Jon Kleinberg, and Éva Tardos. Influential nodes in a diffusion model for social networks. In *Automata, languages and programming*, pages 1127–1138. Springer, 2005. 41

David G Kendall. Branching processes since 1873. *Journal of the London Mathematical Society*, 1(1):385–406, 1966. 22

Ross Kindermann, James Laurie Snell, et al. *Markov random fields and their applications*, volume 1. American Mathematical Society Providence, RI, 1980. 61

Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *Machine learning: ECML 2004*, pages 217–226. Springer, 2004. 82

Gueorgi Kossinets, Jon Kleinberg, and Duncan Watts. The structure of information pathways in a social communication network. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 435–443. ACM, 2008. 92

Dariusz Król. On modelling social propagation phenomenon. In *Intelligent Information and Database Systems*, pages 227–236. Springer, 2014. 40

Frank R Kschischang, Brendan J Frey, and H-A Loeliger. Factor graphs and the sum-product algorithm. *Information Theory, IEEE Transactions on*, 47(2): 498–519, 2001. 67

Bibb Latane. The psychology of social impact. *American psychologist*, 36(4):343, 1981. 35

Kyu-Min Lee, Jung Yeol Kim, Won-kuk Cho, KI Goh, and IM Kim. Correlated multiplexity and connectivity of multiplex random networks. *New Journal of Physics*, 14(3):033027, 2012. 21

Claude Lefevre and Philippe Picard. Distribution of the final extent of a rumour process. *Journal of applied probability*, 31(1):244–249, 1994. 23

Mark R Lepper. Social control processes and the internalization of social values: An attributional perspective. *Social cognition and social development*, pages 294–330, 1983. 35

Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429. ACM, 2007. 63, 73

Cheng-Te Li, Hsun-Ping Hsieh, Shou-De Lin, and Man-Kwan Shan. Finding influential seed successors in social networks. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 557–558. ACM, 2012. 71, 73

Yanhua Li, Wei Chen, Yajun Wang, and Zhi-Li Zhang. Influence diffusion dynamics and influence maximization in social networks with friend and foe relationships. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 657–666. ACM, 2013. 42

David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007. 61

David Liben-Nowell and Jon Kleinberg. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the National Academy of Sciences*, 105(12):4633–4638, 2008. 20

Thomas M Liggett. *Particle Systems*. Springer, 1985. 40

Bo Liu, Gao Cong, Dong Xu, and Yifeng Zeng. Time constrained influence maximization in social networks. In *ICDM*, pages 439–448, 2012. 70

Qi Liu, Biao Xiang, Lei Zhang, Enhong Chen, Chang Tan, and Ji Chen. Linear computation for independent social influence. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 468–477. IEEE, 2013. 66

Qiming Lu, György Korniss, and Bolesław K Szymański. The naming game in social networks: community formation and consensus engineering. *Journal of Economic Interaction and Coordination*, 4(2):221–235, 2009. 42

Matteo Magnani and Luca Rossi. The ml-model for multi-layer social networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 5–12. IEEE, 2011. 21

Vijay Mahajan and Eitan Muller. Timing, diffusion, and substitution of successive generations of technological innovations: The ibm mainframe case. *Technological Forecasting and Social Change*, 51(2):109–132, 1996. 31

Suman Kalyan Maity, Animesh Mukherjee, Francesca Tria, and Vittorio Loreto. Emergence of fast agreement in an overhearing population: The case of the naming game. *EPL (Europhysics Letters)*, 101(6):68004, 2013. 42

Andrei Andreevich Markov. The theory of algorithms. *Trudy Matematicheskogo Instituta im. VA Steklova*, 38:176–189, 1951. 21

Peter V Marsden and Noah E Friedkin. Network studies of social influence. *Sociological Methods & Research*, 22(1):127–151, 1993. 37, 43

Naoki Masuda and Petter Holme. Predicting and controlling infectious disease epidemics using temporal networks. *F1000Prime Reports*, 5:6, 2013. 15, 37, 56, 74

Michael Mathioudakis, Francesco Bonchi, Carlos Castillo, Aristides Gionis, and Antti Ukkonen. Sparsification of influence networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 529–537. ACM, 2011. 69

Clark McCauley. The nature of social influence in groupthink: Compliance and internalization. *Journal of Personality and Social Psychology*, 57(2):250, 1989. 35

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001. 34

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092, 1953. 65

Radosław Michalski, Sebastian Palus, Piotr Bródka, Przemysław Kazienko, and Krzysztof Juszczyszyn. *Modelling social network evolution*. Springer, 2011a. 15

Radosław Michalski, Sebastian Palus, and Przemysław Kazienko. Matching organizational structure and social network extracted from email communication. In *Business Information Systems*, pages 197–206. Springer, 2011b. 81, 92

Radosław Michalski, Piotr Bródka, Przemysław Kazienko, and Krzysztof Juszczyszyn. Quantifying social network dynamics. In *Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on*, pages 69–74. IEEE, 2012a. 15

Radosław Michalski, Jarosław Jankowski, and Przemysław Kazienko. Negative effects of incentivised viral campaigns for activity in social networks. In *Cloud and Green Computing (CGC), 2012 Second International Conference on*, pages 391–398. IEEE, 2012b. 74

Radosław Michalski, Przemysław Kazienko, and Dawid Król. Predicting social network measures using machine learning approach. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 1056–1059. IEEE, 2012c. 61

Radosław Michalski, Przemysław Kazienko, and Jarosław Jankowski. Convince a dozen more and succeed–the influence in multi-layered social networks. In *Signal-Image Technology & Internet-Based Systems (SITIS), 2013 International Conference on*, pages 499–505. IEEE, 2013. 21, 169

Radosław Michalski, Tomasz Kajdanowicz, Piotr Bródka, and Przemysław Kazienko. Seed selection for spread of influence in social networks: Temporal vs. static approach. *New Generation Computing*, 2014. 58, 71, 139, 156

Mauro Mobilia. Commitment versus persuasion in the three-party constrained voter model. *Journal of Statistical Physics*, 151(1-2):69–91, 2013. 42

Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37 (1-2):17–23, 1950. 31

Yamir Moreno, Maziar Nekovee, and Amalio F Pacheco. Dynamics of rumor spreading in complex networks. *Physical Review E*, 69(6):066130, 2004. 23

Gabriel Mugny, Claude Kaiser, Stamos Papastamou, and Juan A Pérez. Intergroup relations, identification and social influence. *British Journal of Social Psychology*, 23(4):317–322, 1984. 35

Katarzyna Musiał, Przemysław Kazienko, and Piotr Bródka. User position measures in social networks. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, page 6. ACM, 2009. 114, 115

Seth A Myers, Chenguang Zhu, and Jure Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–41. ACM, 2012. 19

Maziar Nekovee, Yamir Moreno, G Bianconi, and M Marsili. Theory of rumour spreading in complex social networks. *Physica A: Statistical Mechanics and its Applications*, 374(1):457–470, 2007. 23

John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996. 143

Tore Opsahl and Pietro Panzarasa. Clustering in weighted networks. *Social networks*, 31(2):155–163, 2009. 84

Alberto Palloni. *Theories and models of diffusion in sociology*. Citeseer, 1998. 23

Nishith Pathak, Arindam Banerjee, and Jaideep Srivastava. A generalized linear threshold model for multiple cascades. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 965–970. IEEE, 2010. 40

D Peleg. Local majority voting, small coalitions and controlling monopolies in graphs: A review. In *Proc. of 3rd Colloquium on Structural Information and Communication Complexity*, pages 152–169, 1997. 39

René Pfitzner, Ingo Scholtes, Antonios Garas, Claudio J Tessone, and Frank Schweitzer. Betweenness preference: quantifying correlations in the topological dynamics of temporal networks. *Physical review letters*, 110(19):198701, 2013. 15, 56

Jean Philibert. One and a half century of diffusion: Fick, einstein, before and beyond. *Diffusion Fundamentals*, 2(1):1–10, 2005. 44

Richard R Picard and R Dennis Cook. Cross-validation of regression models. *Journal of the American Statistical Association*, 79(387):575–583, 1984. 144

Christina Prell. *Social network analysis: History, theory and methodology*. Sage, 2011. 9, 31

Singiresu S Rao and Vimal Singh. Optimization. *Systems, Man and Cybernetics, IEEE Transactions on*, 9(8):447–447, 1979. 71

Bertram H Raven. Social influence and power. Technical report, DTIC Document, 1964. 33

Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70. ACM, 2002. 51, 54, 55, 61

Garry Robins, Philippa Pattison, and Peter Elliott. Network models for social influence processes. *Psychometrika*, 66(2):161–189, 2001. 37, 43

Everett M Rogers. *Diffusion of innovations*. Simon and Schuster, 2010. 24, 25, 26, 27, 28, 29

Tim Rogers and Thilo Gross. Consensus time and conformity in the adaptive voter model. *Physical Review E*, 88(3):030102, 2013. 42

Richard M Ryan and Jerome Stiller. The social contexts of internalization: Parent and teacher influences on autonomy, motivation, and learning. *Advances in motivation and achievement*, 7:115–149, 1991. 35

Stanisław Saganowski, Piotr Bródka, and Przemysław Kazienko. Influence of the dynamic social network timeframe type and size on the group evolution discovery. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 679–683. IEEE, 2012. 12

Stanisław Saganowski, Piotr Bródka, Anna Zygmunt, Przemysław Kazienko, and Jarosław Koźlak. Predicting community evolution in social networks. *AI Communications, in reviews*, 2014. 12

Kazumi Saito, Ryohei Nakano, and Masahiro Kimura. Prediction of information diffusion probabilities for independent cascade model. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 67–75. Springer, 2008. 42, 67

Thomas Schelling. *Micromotives and Macrobehavior*. WW Norton and Company, New York and London, 1978. 38

Alexander Schrijver. *Combinatorial optimization: polyhedra and efficiency*, volume 24. Springer, 2003. 62

Paulo Shakarian and Damon Paulo. Large social networks can be targeted for viral marketing with small seed sets. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 1–8. IEEE Computer Society, 2012. 65

Jimeng Sun and Jie Tang. A survey of models and algorithms for social influence analysis. In *Social network data analytics*, pages 177–214. Springer, 2011. 37

Taro Takaguchi, Naoki Masuda, and Petter Holme. Bursty communication patterns facilitate spreading in a threshold-based epidemic dynamics. *PloS one*, 8 (7):e68629, 2013. 74

Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816. ACM, 2009. 67, 75

Lei Tang, Huan Liu, Jianping Zhang, and Zohreh Nazeri. Community evolution in dynamic multi-mode networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 677–685. ACM, 2008. 92

RDevelopment Core Team et al. R: A language and environment for statistical computing. *R foundation for Statistical Computing*, 2005. 102

Douglas Tigert and Behrooz Farivar. The bass new product growth model: a sensitivity analysis for a high technology product. *The Journal of Marketing*, pages 81–90, 1981. 31

Jeffrey Travers and Stanley Milgram. An experimental study of the small world problem. *Sociometry*, 32(4):425–443, 1969. 20

Daniel Trpevski, Wallace KS Tang, and Ljupco Kocarev. Model for rumor spreading over networks. *Physical Review E*, 81(5):056102, 2010. 23

Fang-Mei Tseng and Yi-Chung Hu. Quadratic-interval bass model for new product sales diffusion. *Expert Systems with Applications*, 36(4):8496–8502, 2009. 31

John C Turner. *Social influence*. Thomson Brooks/Cole Publishing Co, 1991. 34

Thomas W Valente. Network models of the diffusion of innovations. *Computational & Mathematical Organization Theory*, 2(2):163–164, 1996a. 31

Thomas W Valente. Social network thresholds in the diffusion of innovations. *Social networks*, 18(1):69–89, 1996b. 31

N Venkatraman, Lawrence Loh, and Jeongsuk Koh. The adoption of corporate governance mechanisms: a test of competing diffusion models. *Management Science*, 40(4):496–507, 1994. 33

Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42. ACM, 2009. 87

Stanley Wasserman and Katherine Faust. Social network analysis: Methods and applications. 1994. 9, 81, 114, 115

Duncan J Watts and Steven H Strogatz. Collective dynamics of small-world networks. *nature*, 393(6684):440–442, 1998. 20, 113, 115

Malcolm Wright, Clinton Upritchard, and Tony Lewis. A validation of the bass new product diffusion model in new zealand. *MARKETING BULLETIN-DEPARTMENT OF MARKETING MASSEY UNIVERSITY*, 8:15–29, 1997. 31

Jierui Xie, Sameet Sreenivasan, György Korniss, W Zhang, C Lim, and Bolesław K Szymański. Social consensus through the influence of committed minorities. *Physical Review E*, 84(1):011130, 2011. 42

Jierui Xie, Jeffrey Emenheiser, Matthew Kirby, Sameet Sreenivasan, Bolesław K Szymański, and György Korniss. Evolution of opinions on social networks in the presence of competing committed groups. *PloS one*, 7(3):e33215, 2012. 42

Bo Xu and Lu Liu. Information diffusion through online social networks. In *Emergency Management and Management Sciences (ICEMMS), 2010 IEEE International Conference on*, pages 53–56. IEEE, 2010. 23

Jaewon Yang and Jure Leskovec. Modeling information diffusion in implicit networks. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 599–608. IEEE, 2010. 21, 23

Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social Media Mining: An Introduction*. Cambridge University Press, 2014. 39, 41

Anita Zbieg, Blazej Zak, Jarosław Jankowski, Radosław Michalski, and Sylwia Ciuberek. Studying diffusion of viral content at dyadic level. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis*

and Mining (ASONAM 2012), pages 1259–1265. IEEE Computer Society, 2012. 23

Weituo Zhang, Chjan Lim, György Korniss, and Bolesław K Szymański. Spatial propagation of opinion dynamics: Naming game on random geographic graph. *arXiv preprint arXiv:1401.0115*, 2013. 42