Philipp Koncar, BSc

Analyzing Dynamical Aspects of Activity in Social Networks

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme: Software Development and Business Management

submitted to

Graz University of Technology

Supervisor

Assoc.Prof. Dipl.-Ing. Dr.techn. Denis Helic

Knowledge Technologies Institute

Graz, October 2015

Philipp Koncar, BSc

Eine Analyse der dynamischen Aspekte von Aktivität in sozialen Netzwerken

MASTERARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

Masterstudium Softwareentwicklung - Wirtschaft

eingereicht an der

Technischen Universität Graz

Betreuer

Assoc.Prof. Dipl.-Ing. Dr.techn. Denis Helic

Institut für Wissenstechnologien

Graz, Oktober 2015

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all ma- terial which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis dissertation.

Graz, _____ _____

Date Signature

Eidesstattliche Erklärung[1]

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

Graz, am _____ _____

Datum Unterschrift

[1] Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

Acknowledgements

First, I want to thank my supervisor, Prof. Denis Helic, for the amazing opportunity to write this master's thesis and to work with his team at the Knowledge Technology Institute at Graz University of Technology. Furthermore, I want to thank Simon Walk for all his great work, never-ending support and answering my countless questions during the last year. Your capabilities in teaching are truly inspiring and made this whole experience very enjoyable.

I also want to thank the numerous acquaintances that I have made during my studies, especially Christof, David and Markus for all the assignments we have done and all the fun we have had together.

Last but certainly not least, I want to thank my family and close friends for always supporting me in what I am doing, no matter how moronic it may appear to them.

The successful completion of my studies would have never been possible without all of you, so my sincere gratitude goes to every single one of you.

Graz, October 7, 2015 Philipp Koncar

Abstract

One determinant of success for websites on the Internet is activity of their users. This especially holds for online collaboration networks, where users join lively and highly dynamic communities and engage in various collaborative activities and interactions. An example for such a network is the StackExchange.com web portal, providing its users the possibility of asking questions, answering them, as well as commenting on them and voting for them to express their relevance. Usually, the evolution of such networks is influenced by various external and internal factors. For example, new users join; existing ones leave; new collaborations between existing users arise or are discontinued for whatever reason. Hence, website owners are in need for aligned tools and models to understand user behavior, eventually enabling them to better predict, prevent and cope with the consequences of such occurrences.

In this thesis the existing *Activity Dynamics* framework—based upon the principles of dynamical systems on networks and used for simulating activity—is adopted and extended to facilitate three different experiments. First, we conduct six plausible real-world scenarios to analyze their impacts on simulated activity for ten different sized empirical data sets—five StackExchange.com and five Semantic MediaWiki networks. Second, we use these data sets to analyze the correlations between user centrality within the network and simulated activity. Third, we further extend the *Activity Dynamics* framework by removing restrictions of the underlying static network structure and introduce a dynamic version of the model in order to increase the accuracy of the simulations.

Kurzfassung

Benutzeraktivität ist ein Erfolgsfaktor für Internetseiten im World Wide Web. Dies gilt vor allem für Online-Kollaborationsnetzwerke, in denen Benutzergemeinschaften zusammenarbeiten und interagieren. Ein Beispiel für ein Netzwerk dieser Art ist StackExchange.com, in dem Benutzer Fragen stellen und beantworten können, die Möglichkeit haben Fragen und Antworten zu kommentieren und ihre jeweilige Relevanz über Bewertungen angeben können. Die Aktivität in solchen Netzwerken kann von unterschiedlichen internen und externen Faktoren beeinflusst werden. Zum Beispiel können neue Benutzer dem Netzwerk beitreten, bestehende Benutzer es verlassen, neue Kollaborationen zwischen Benutzern entstehen oder sich vorhandene auflösen. Neue Ansätze und Tools sind nötig, um die Besitzer und Administratoren von Online-Kollaborationsnetzwerken im Umgang mit diesen Situation unterstützen zu können.

In dieser Masterarbeit wird das bestehende *Activity Dynamics* Framework, das auf den Prinzipien von dynamischen Systemen beruht und für die Simulation von Aktivität in Online-Kollaborationsnetzwerken verwendet wird, adaptiert und erweitert um drei unterschiedliche Experimente durchführen zu können. Zuerst werden sechs plausible Szenarien vorgestellt, um deren Einfluss auf zehn empirische Datensätze – fünf StackExchange.com- und fünf Semantic MediaWiki-Netzwerke – zu analysieren. Des Weiteren werden diese Datensätze für die Analyse von Korrelationen zwischen Benutzerzentralität im Netwerk und der simulierten Aktivität verwendet. Schließlich wird das *Activity Dynamics* Framework weiter ausgebaut, um Einschränkungen der zugrundeliegenden statischen Netzwerkstruktur aufzuheben und die Genauigkeit der Aktivitätssimulation zu erhöhen.

Contents

List of Figures

List of Tables

1 Introduction

In this thesis we deal with the simulation and analysis of activity dynamics in social online communities. The results provided and analyzed in this work build upon previous work done by Walk et al. (2015) where they developed the *Activity Dynamics* framework, which is briefly explained in Section 3.1, and is used for the experiments conducted in this work.

1.1 Motivation

The success of any website strongly depends on its content and the relevance to its visitors. In online collaborations networks the content is usually generated by their participants and therefore the activities of those collaborators play an important part. An example of such a network is the StackExchange.com website, a question and answering portal where users can post questions, answer existing questions of other users, comment on answers and questions and distinguish relevant topics by up or down voting them. Another example of an online collaboration network are online encyclopedias, such as Wikipedia, where authors create and edit articles in order to provide various information to their readers. It seems obvious that without rich user participation, the collaboration network may die out if no action is taken, thus no new content is provided and the website may not acquire target visitor numbers. This is the worst case for any advertising-financed website as the cash flow stops flowing. To better understand their users and the overall activity they create, new technologies and tools are needed. Additionally, such tools could support website owners in analyzing the kinds of influences that potentially cause activity within a network to increase or decrease.

Since the activity in collaboration networks can be affected not only by intrinsic factors but also by many external factors, it is hard to predict the results of intended or unforeseen events that occur during a network's life cycle. For example, the introduction of new terms-of-service with a change in the privacy policies or successful attacks of hackers exposing security issues within the platform, could lead to a massive drop in user numbers and harm overall activity. On the other hand, the introduction of a new feature to an existing system or a sales promotion could lead to an extraordinary gain in users, resulting in a sudden increase of overall activity that could overload the servers of networks. Hence, website owners and administrators are in need of tools that provide information about the possible outcomes whenever such events are talking place, either if they are intended or unexpected. This would allow them to prepare early and take preventive actions that are best for the network's overall activity.

One possible way to implement such supporting tools is using the *Activity Dynamics* framework. It is based on the principles of dynamical systems on networks and can be configured with one single parameter. In addition, it is fairly easy to manipulate and to extend its underlying source code. This grants us the abilities to model and simulate activity dynamics and activity trends for online collaboration networks. We illustrate the accurate simulation performance by comparing the activity simulated with the help of this framework and empirical data gained from real life online collaboration networks. With this in mind, we uncover new possibilities of providing answers about the impact of various events.

1.2 Objectives

This master's thesis covers the following objectives:

(i) To briefly introduce the reader to the *Activity Dynamics* framework by explaining its mechanisms, variables and setup parameters.
(ii) To identify and simulate six plausible real world scenarios for ten different real word collaboration networks to uncover the resulting implications.

(iii) To discover and analyze correlations between the centrality measures of users within the network and the simulated activity.
(iv) To improve the *Activity Dynamics* framework by considering the changes in the underlying network structure over time in order to increase the simulation accuracy.

1.3 Contributions

This work tries to gain insights in the dynamics of activity in online collaboration networks. Based on the obtained results, we can make general assumptions about the impacts of plausible real world events and the importance of structural influences within such networks. For that reason, we conduct three different experiments:

Activity Dynamics Scenarios. In this experiment we define six scenarios that—combined with simulations of the *Activity Dynamics* framework—allow for hypothetical explanations of potential events talking place in real-world online collaboration networks. Beside the aforementioned events, where user numbers either decline or rise, four other possible events are taken into account: The sudden stop of interactions between existing users which results in the removal of collaboration ties, as well as the possibility of existing users starting to socialize and therefore create new collaboration ties. In another scenario we introduce external incentives to collaboration networks. Such incentives could, for example, be paid moderators supervising collaborations, aiming to increase activity. Furthermore, we investigate the occurrence of trolls—users trying to harm the network—and their impact on the overall activity. We conduct all scenarios in two different approaches: By either *randomly* selecting affected users or by specifically selecting them based on their importance, which represents our *informed* approach. Finally, we apply this experiment to five StackExchange.com networks and to five Semantic MediaWikis—each represented by graphs where users are nodes and collaboration ties are edges—to show the impacts of our scenarios. Table 1.1 provides a quick overview of all scenarios and explains the simulated events occurring in our empirical networks.

Table 1.1: **Summary of the Investigated Scenarios Conducted as Part of This Thesis.** The *Scenario* column states the name for each scenario. *Objectives* describe the different events we simulate for each empirical network.

Scenario	Objective
Mass Emigration	Users seize contribution in the collaboration networks and leave.
Mass Immigration	Incidents attract new users that join the networks abruptly.
Breaking Collaborative Ties	Users lose interest in each other and decide to not longer collaborate.
Establishing New Collaborations	A recommendation system causes new collaboration edges between existing users.
Providing Incentives	Providing incentives to users in order to raise the overall activity.
Emergence of Trolls	Newly emerging trolls influence the activity in collaboration networks.

Centrality Analysis. Further, the *Activity Dynamics* framework allows us to investigate the role of collaboration network topologies. The topology of a network is on the one hand its spectrum—the eigenvalues and eigenvectors of the adjacency matrix of the representing graph—and on the other hand the structure of the network. According to our model, we show that the underlying structure of the graph correlates with simulated activity and thus allows us to identify most active and important users in the collaboration network by simply considering a user's centrality. We measure this centrality in three different ways: By the degree (number of connected neighbors of a user), the eigenvector centrality and Google's PageRank. Further, we compare the resulting correlations between simulated activity and user centrality by calculating correlation coefficients.

Dynamic Network Structure. The original *Activity Dynamics* framework might not be as accurate as it could be due to the fact that it uses a static network structure throughout the whole simulation process and its parameters are only calculated once for the whole network and time span. Hence, we further extend it, enabling it to take changes in the network structure over time into account. This allows us to calculate parameters for all points in time (for example per month) and therefore to increase the simulation performance, making it even more reliable and useful for website owners and administrators.

The conducted experiments and their results aim to provide new tools for website owners, which help them to better understand user behavior within their online collaboration networks. We accomplish this with the help of previous work done by Walk et al. (2015), where the authors have

introduced the *Activity Dynamics* framework, which is used to simulate activity and serves us as a skeletal structure for our experiments. With this framework we can preprocess the empirical data sets, initialize the model, calculate its parameters and simulate activity for the desired time span. However, various extensions had to be implemented in order to simulate the presented scenarios, to calculate correlations, to achieve the dynamic version of the *Activity Dynamics* framework and to obtain useful results from these experiments. The changes to the original version are described from Chapter 4 to Chapter 6.

1.4 Thesis Outline

This work is structured into 9 chapters. Chapter 2 follows the introduction and provides an overview of the related work in the fields of dynamical systems, epidemic models and opinion dynamics. In Chapter 3 we explain the used empirical data sets in detail and describe the *Activity Dynamics* framework, as well as the calculation of parameters needed for simulations. In Chapter 4 we depict the six scenarios and their different approaches conducted as part of this work. In Chapter 5 we present the correlation analysis between simulated activity and three different centrality measures. Chapter 6 lists the changes made to the original *Activity Dynamics* model in order to use a dynamic network structure during simulations and depicts how the calculation of parameters has changed compared to the original *Activity Dynamics* framework. The results of our three different experiments are described in Chapter 7. This contains the impact of the six scenarios on our empirical data sets, the calculated correlations between user centrality and simulated activity, as well as the outcome of using a dynamical network structure. We interpret these results in Chapter 8. In Chapter 9 we conclude this master's thesis and present future work.

2 Related Work

All results depicted in this thesis draw upon the theories and principles of *dynamical systems* and *epidemic models* on networks explained in Section 2.2 and in Section 2.3.

2.1 Background

The evolution of social group dynamics or social movements have been observed several times in previous research. For example, Milgram (1967) showed in his experiments that information—given the right conditions—can spread quickly through real-world social networks. Other works investigated peer pressure and explained the intentions of individuals joining groups as, for example, in Waddington and Whitston (1997) and Putnam, Leonardi, and Nanetti (1994). Stark and Bainbridge (1980) explained the importance of interpersonal bonds between existing members and future group members when it comes to group growth. However, these results are all built solely upon empirical studies. Since the emergence of online social networks, new ways to study the fields of group dynamics have emerged and have since been used to increase our understanding of the complex processes that occur in such systems.

Critical Mass Theory. The first concepts of research in the field of online social communities rely on the formal theory of collective action and critical mass. In Physics, the critical mass is the amount of radioactive material needed in order to reach nuclear fission. From a sociological point of view, this term is used in a metaphorical way, indicating that a certain threshold of user numbers or activity needs to be reached before a social movement starts evolving. In 1985, a first explanation of critical mass, in the context of social

networks, was made by Oliver, Marwell, and Teixeira (1985). Corresponding to them, the starting point of individuals forming a group and its evolution of activity is based on two independent variables: The *production function* and the *group heterogeneity*. The first variable indicates the expected outcome of an individual's efforts taken in a collaborative environment. In other words, it outlines the return of the contribution done on different levels of effort. Most likely shapes of production functions are: accelerating, decelerating or linear. Given an accelerating production function, it means that early efforts made by an individual have minimal effect in the beginning, but pay off more effectively later on. In contrast to that is the decelerating production function, where efforts made in the beginning have a bigger effect on the collective good, but benefits begin to shrink in later phases. With the linear production function, each contribution has a similar impact on the value of collective goods.

The second variable mentioned in this work is the group heterogeneity. It describes different interests of individuals in certain situations. For example, the construction of a new gas station might be more important for car owners than for individuals not having any fuel powered machines. This could lead to individuals that do not contribute to the collective good, but still benefit from it. This circumstance was already discussed by Olson Jr (1965), where he described the exploitation of the ones contributing more to the collective good by the ones that do not participate at all. In addition to that, they found that groups with higher heterogeneity are more likely to produce an accelerating production function.

When putting this theory in the context of online social communities, the collective good can, for example, be seen as a topic all community members are interested in. However, in these situations it is not quite clear if the principles of critical mass theory still apply. One problem with critical mass theory might be that it is sequential, meaning each individual is influenced by those that acted before them as described by Markus (1987). More realistically, users are usually influenced in both ways. If, for example, user A posts a question on one of the StackExchange.com question and answering portals and gets an answer from another user B who is participating in that network as well, it can influence user A in his or her further reactions, such as posting a comment on that answer or posting a follow-up question.

Solomon and Wash (2014) have used this approach to explore the community growth in WikiProjects. WikiProjects are groups of Wikipedia authors who work together in order to improve the online encyclopedia Wikipedia. Hence, users in such a network are typically organized around one topic (or one collective good) or at least around topics with similar context. In their work, they tried to answer what critical mass in an online social community could look like. Basically, they tried to find out what such a network needs to become self-sustainable, meaning the network will theoretically stay active forever without the need of external influences. Their findings show that when a few users are submitting large amounts of content in early stages, it will equally result in lower growth rates. Better long-term growth can be reached when many different individuals participate in small amounts. This led the authors to the assumption that online communities are more sustainable when they grow by content created by newly joined users and not by a few long term participants. This fact is even more crucial when new users have a high diversity in interests.

Based on this very basic concept, different approaches and models can be developed to model activity in dynamical social systems more precisely. Nevertheless, one question that already can be answered with this idea and the knowledge gained in this simple variant of activity simulation, is how many individuals need to participate in an online collaboration network in order to raise activity and probably gain self-sustainability. However, to gain even more knowledge of the various dynamics in such networks, a more detailed approach is needed.

The *Activity Dynamics* framework, presented and used in this work, is based on the idea of critical mass theory but extends this concept by using the principles of dynamical systems in order to gain higher levels of accuracy.

2.2 Dynamical Systems

In general, a dynamical system is a mathematical model describing a physical problem. Mathematicians and physicists have various definitions for it. Katok and Hasselblatt (1997) stated a most general notion, in which dynamical systems consist of a *phase (or state) space* in which an arbitrary

amount of elements represent possible states of the system. As time evolves, the systems encounters state transitions. Each state x of the system depends on time t, that may either be discrete or continuous, with the possibility of representing not only future evolution of a system, but also states of the system in the past. For example, time may be an integer variable for a discrete dynamical system or a real variable for a continuous dynamical system. In addition to these two principles, the *time-evolution law* specifies how a state in the next period of time depends on the initial input of the system and on the state in the previous period of time. Furthermore, Strogatz (2001) explained the different *attractors* a dynamical system moves toward during evolution of time. First of all, the *point attractor* where a non-linear dynamical system at a certain point $x(t)$ may come to a rest. Then the velocity at that point must be zero so this point is called a *fixed point* denoted by x^*. Whenever a system reaches this fixed point, all state transitions stop and the system stays in its currently prevailing state forever. This state is also called the *equilibrium*. Another possibility is that the state of a system circulates around a closed loop of states forever. Whenever the system reaches this set of attractors, called *limit cycle*, it represents an oscillation of the dynamical system. A third possibility would be that the system is attracted by a *strange attractor* where it wanders forever and never stops. This occurrence results in an unstable or chaotic behavior of the dynamical system. A more in-depth introduction to dynamical systems can be found in Strogatz (2014), Barrat, Barthelemy, and Vespignani (2008) and in Newman (2010).

2.2.1 Dynamical Systems on Networks

Newman (2010) explained how the principles of dynamical system can be applied to networks. Here, each vertex i of a network has a set of independent dynamical variables $x_i, y_i, \ldots$ and is connected to other vertices over edges. According to Newman and given a dynamical system with a single variable x, the time evolution of x_i is denoted by:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = f_i(x_i) + \sum_j A_{ij} g_{ij}(x_i, x_j), \tag{2.1}$$

where f_i states the intrinsic dynamics of vertex i and g_{ij} states the concurrence of vertices connected to i. A is the adjacency matrix of the network, where $A_{ij} = 1$ if i and j are connected and $A_{ij} = 0$ otherwise. Note that there are different functions f_i and g_{ij} for each unique vertex. However, as each vertex can represent the same property, the dynamics of all vertices can also be the same. Hence, Equation 2.1 is further simplified to:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = f(x_i) + \sum_j A_{ij} g(x_i, x_j), \tag{2.2}$$

making both terms, f and g, the same for each vertex within the network. The *Activity Dynamics* framework uses Equation 2.2 to model activity of users in online collaboration networks denoted by variable a. We will further describe this in Section 3.1.

2.2.2 Diffusion Processes on Networks

Dynamical systems have been used to model different economical and social processes taking place in social networks. The main focus was set on information diffusion processes, which try to model how information is spread in online and offline social communities.

The following presented models all take different approaches to simulate diffusion of information or to find the most influential users in such communities. Various approaches build upon *Epidemic Models*, a subcategory of dynamical systems and further described in section 2.3. For example, Leskovec, L. A. Adamic, and Huberman (2007) used an epidemic approach to identify the dynamics of viral marketing. By analyzing a recommendation network, they found that individuals with a high degree play a very important part in such networks. However, their models assume that high degree users have as much probability of influencing each of their neighbors as the individuals with low degree do. They found that there are limits to how much influence the important users have in a social network, suggesting that each user only influences a few of his neighbors and not all users within the network. Finally, they stated that the spreading of information in the sense of viral marketing is not as epidemic as marketers have hoped.

Another application of dynamical systems on networks was made by Centola and M. Macy (2007). In their work they explained the strength of weak ties. Weak ties—in a structural meaning—are connections in graphs that connect different distant components with each other and can therefore be also called long ties as suggested by Granovetter (1973). They found, that for random graphs only a few of such weak ties are needed in order to spread information in social networks. However, for structured networks resulting from social interactions between real users, this might not always be the case. Using Watts and Strogatz's original model described in D. J. Watts and Strogatz (1998), they stated that long ties can also be a disadvantage for the spreading of information in such networks. In fact, only long and narrow ties can be useful for information diffusion, but as soon as too much randomness is added, spreading of information might become inefficient.

A very simple and basic model is explained by D. J. Watts (2002), where the author tried to explain how social networks are effected by *cascades*. Cascades describe social phenomena in which a single action taken by one individual results in a wave of actions taken by other participants of the same network. The author used a random network where each vertex represents an agent and each agent's decisions are determined by the decisions of their neighbors. He discovered a simple binary-decision model which allows for setting up testable predictions about cascades in real social systems. When the social network is sparse, the propagation of a global cascade is limited by the global connectivity of the network. However, when it is dense, the propagation of a cascade is limited by the local stability—the more neighbors they have, the more stable they are—of nodes. Furthermore, the work showed that increased heterogeneity of individuals appears to increase the likelihood of global cascades. However, high heterogeneity of the degree of individuals appears to reduce this probability.

Manuel Gomez Rodriguez, Leskovec, and Krause (2010) introduced *NetInf*, an algorithm to infer networks of information diffusion. It enabled them to study properties of real-world networks. The model was evaluated on large real-world data sets of memes spreading on various news websites. They discovered that clusters of websites with similar topics are able to diffuse information from one cluster to another due to a rather small number of social hubs connecting these different clusters.

The authors build upon their previous work and developed the *NetRate* model in M Gomez Rodriguez et al. (2011). It infers transmission rates between individuals of a social network. By introducing continuous temporal dynamics, they avoided further needed assumptions and simplified problems existing in their older model explained in Manuel Gomez Rodriguez, Leskovec, and Krause (2010). This improved model uses parameters with natural interpretations, leading to a maximum likelihood problem that can be solved efficiently. The parameters do not require any manual fine tuning and are calculated automatically based on empirical data.

Adar, L. Adamic, et al. (2005) studied the propagation of information and memes on weblogs, also referred to as blogs. In their work they presented a system that allows for visualization of information flow between such weblogs. The main interest lies on the path that information takes while it is propagating through the world wide web. Their work is related to *link inference* (the problem of inferring types of links), for example explained and used in Berger and Bommel (1996) and Aggarwal, Xie, and Philip (2012), and *link classification* (classification based on link structure of networks), for example explained and used in Lu and Getoor (2003). In addition to that, they used non-traditional features that are unique for blogs, facilitating better results than previous models. However, they also stated that incompleteness of crawling through the web may lead to errors. Another possible drawback in their approach is the problem of memes and information represented by different URLs, for example one unique image hosted on different providers. With this in mind, paths of information flows may not be discovered completely by their model.

Kempe, Kleinberg, and Tardos (2003) introduced the *Cascade Model for Information and Knowledge Diffusions* and showed how diffusion of information in social networks can be maximized. Their results are important for marketing engineers and show how peer influence and word-of-mouth effects are an important factor in the dynamics of information diffusion.

Another model was introduced by Goyal, Bonchi, and Lakshmanan (2010), where they found a way to investigate probabilities of interactions that are represented by weights of edges in an influence propagation network. By using logs from past propagations with static and time-dependent models for calculating these probabilities of influence, their algorithms are able to

predict whether a user will engage in an action or not. These algorithms are performing extremely well and accurate on users with high influence. In addition to that, the algorithms are able to more or less predict the point in time at when a user might engage in an action. Furthermore, the static approach described in this paper performs very close to the more complex and more run-time intensive time-dependent variant.

Tang et al. (2009) introduced a solution to topic-based social influence analysis. In their work, they suggested a *Topical Affinity Propagation* approach to describe these analyses by using a graphical probabilistic model. Their learning algorithm is based on a map-reduce programming model which solves the efficiency problem that occurs with run-time of simulation. In addition to that, the algorithm scales in an advantageous way. This algorithm turned out to improve the performance of expert finding in social networks, which is still an open problem.

Cha et al. (2010) analyzed the influence of twitter users on each other. For that, they considered three different factors: the number of followers, the number of retweets and the number of mentions. They found, that the number of followers is just representing the popularity of a user but not necessarily the influence this user has on the rest of the network's users. Additionally, they showed that decisive user influence—that is positive or negative influence on other users—is not gained spontaneously, but with great personal involvement and effort. They stated that influential users in a network are indeed able to be located, which is in contrast to previous claims made by D. Watts (2007), where he stated that influential users of a network are hard to identify.

In Lappas et al. (2010), the authors tried to find *effectors*, a set of users that lead from an initial state of activity to an observed finite state of activity of a network which is represented as a graph. In their work, they described that the problem of finding the belonging effectors that initially started activity propagation are NP-hard to solve or even to approximate. However, they found that the problem can be solved in polynomial time for directed trees. Hence, they offer a framework, working with a directed influence graph and an activation vector. They can first find a tree that spans all active users in a network, and then, by using dynamic programming, they could find

the optimal set of effectors in this tree. The results for this approach turned out to perform well and accurate compared to real-world empirical data.

Furthermore, Cosley et al. (2010) compared two models dealing with the measurement of user influence—one based on observed *snapshots* of empirical data and the other one based on continuous temporal dynamics—in order to discover a possible relationship between them. Using data from the English, German and French version of Wikipedia, they showed that approximations of user influence done with the snapshots is not as close to the dynamic version as one would expect, but still useful for comparison. This may help to better understand user influence by allowing for easier comparison of data gathered with different kinds of sampling.

The *Activity Dynamics* framework is based on dynamical systems but sets the focus on activity dynamics of online collaboration networks in its entirety. Users—represented by nodes in a graph—are initialized with activity weights that further depend on intrinsic user behavior and external influences of neighbors. With continuous time, these node weights evolve and simulate empirically observed levels of activity.

2.3 Epidemic Models

A subcategory of dynamical systems are the so-called epidemic models. Initially, epidemic models were developed to learn about and explore the spreading of diseases. Gaining knowledge about the characteristics of highly contagious diseases has been an issue for scientists for a long time. The first assumptions about implications of outbreaks have been made by W. O. Kermack and A. G. McKendrick (1927). Ever since, different ways to explain the outbreak and spreading of certain diseases, such as HIV or smallpox, have been developed. These models aim to help in designing prevention plans and modeling effects of vaccinations. There are various versions of epidemic models but usually they consist of a set of possible states each individual can adopt and the probabilities of transitions between those states. The fact if an outbreak becomes an epidemic is depending on the *epidemic threshold*, for example explained closer by W. Kermack and A. McKendrick (1932). This threshold is the minimum density of infected

individuals needed at the beginning of simulations in order to potentially infect all other individuals within the network over time.

The simplest model among them is the **SI model**, where individuals can only obtain two different states, either *susceptible* in which they do not have a disease or—the other way around—*infected* in which they do have a disease and are able to pass the disease to susceptible individuals. At every unit time and for each individual there is an average number of contacts that are made with randomly chosen other persons. The rate of change from susceptible to infected is modeled with a simple differential equation depending on infection probability β. In addition to that, Individuals that became infected will stay in this state forever. Besides the SI model, there are further models that introduce new states, such as the **SIR model** (susceptible - infected - recovered), in which individuals are able to recover from a disease and are immune after that, so basically they can not become infected anymore. This happens with the probability γ, stating how long infected individuals are staying in the infected sate before the recover and reach the recovered state. Another model is the **SIS Model** (susceptible - infected - susceptible), where susceptible individuals can become infected and can afterwards recover but are not immune to the disease they got in the first place. Furthermore, this model is extended by the **SIRS model** (susceptible - infected - recovered - susceptible) model, where individuals are immune once they recovered, but still have a chance to become susceptible for a new infection again. For more information on epidemic models, the interested reader is pointed to the work of Pastor-Satorras et al. (2014) and of Newman (2010).

While the first epidemic models were used to simulate the outbreak of diseases (for example May and Anderson (1984), Yorke, Hethcote, and Nold (1978), and Lloyd and May (1996)), its principles have been used for other diffusion processes, for example information diffusion, activity dynamics or opinion dynamics. In the past there have been numerous experiments that try to model non-linear social systems and the traditional epidemic models (SI, SIR, SIS and SIRS) have since been further developed to be able to consider the underlying network structure. The following paragraphs cover an overview of these experiments and their various underlying models.

Rvachev and Longini (1985) presented a model that is formulated with difference equations in a continuous state space and discrete time domain

in order to predict the spreading of the influenza virus. They found that their model could successfully predict the geographic spread of Hong Kong's influenza pandemic in 1968. Furthermore, it is considered to be a milestone in the history of epidemic models since it was the first time someone simulated an epidemic outbreak for a period of 425 days.

Another approach is explained by Ferguson, Keeling, et al. (2003), where they tried to extend the original SIR epidemic model by considering various other factors to predict the outbreak of smallpox. They did that by capturing the social and spatial structure and introduced them into simulations. One possibility to do so is taking the underlying network structure into account. They described different types of network structures, such as families, friends and working colleagues, where possibility of infection is depending on the structure of the underlying contact network. They uncovered a considerably large amount of uncertainties that led to the final assumption that no model can truly predict the spreading of smallpox. Hence, they suggested that modeling should aim the identification of effective interference for a variety of outbreak scenarios.

Hufnagel, Brockmann, and Geisel (2004) used an extended SIR model to simulate outbreaks of diseases on a global scale. They used a real-world global aviation network, where airports are nodes and flight paths are represented by edges. Edge weights are given by the number of passengers traveling a flight path per day. In their results, they explained that isolating the largest cities is more effective than removing the edges with highest weights from the network. This suggests that removing most important nodes will harm a network more than removing the most weighted edges.

Similar to this, Ferguson, Cummings, et al. (2005) also took network structure into account and used data from the international air transport association to construct a network based on real-world data. Again, nodes represent airports and edges represent flight paths weighted with the amount of passengers taking this route. They came to the result that it is highly important to bring in more details in the dynamics of disease outbreaks in order to gain more accurate results appropriate to the real world.

Longini et al. (2005) extended the original SIR model to stem the pandemic influenza and prevent an outbreak. In their work, they constructed a network consisting of different sized clusters where each cluster has other

probabilities of infection. They found that an influenza could be prevented from spreading within the first 21 days after the outbreak. However, this is strongly depending on the reproduction number, stating how fast individuals are infected.

Another application of epidemic models is the simulation of computer viruses spreading through networks. Kephart and White (1991) used the SIS model on random graphs to simulate such outbreaks. Using deterministic approximation, stochastic approximation and simulation, they obtained results that were essentially identical to the classical homogeneous theory: They found that epidemics can not spread until the rate of an individual infecting a susceptible individual exceeds the rate at which an individual leaves the infected state and becomes susceptible again. If, however, the former rate is higher, the epidemic is more likely to happen. The number of infected individuals in equilibrium can only be zero or all individuals that are part of the network, depending on this rate. With their simulations they showed that these theoretical results are true as long as the network is highly connected. However, if the network is loosely connected, the epidemic threshold is highly increased. In another simulation they added weak ties to a random graph, resulting in an increased epidemic threshold. The effects of locality were simulated in spatial model simulations, revealing a polynomial growth rate of infections in contrast to the exponential growth rate of the random graphs model.

An interesting model based on the principles of epidemic models and to simulate the spread of computer viruses, was developed by Kephart, Sorkin, et al. (1997). In their work, they mentioned that the basic epidemic model is inadequate and needs adjustment in order to better reflect reality. The error lies in the assumption of uniform chances of infection between every individual in the whole population, suggesting that further parameters are needed to bring simulation results closer to reality.

Kephart and White (1993) introduced two more extended epidemic models in order to explain phenomena occurring in the fields of computer viruses. One of them is the extension of the SIS model by introducing kill signals. Basically this allows healing of infected systems to be dependent on other individuals in the network. For example, if one individual knows about an infected system, he can warn others to take prevention actions. Thus, curing

of infected individuals can be influenced externally. This was one of the first models capable of considering this circumstance.

Furthermore, Y. Wang et al. (2003) developed a more precise model to study the spreading of computer viruses, demonstrated by the application in synthetic and real-world networks. Using only a single parameter—the highest eigenvalue of the adjacency matrix of the graph—their epidemic threshold is more precise than in previously developed models and studies. Furthermore, they depicted that whenever the infection rate is below the epidemic threshold, the amount of infected individuals increases exponentially over time.

The importance of network topology and network spectra (eigenvalues of the adjacency matrix of a graph) has been further discussed by Ganesh, Massoulié, and Towsley (2005). They developed conditions in which the outbreak of an epidemic either dies out quickly or slowly. These conditions hold for random graphs, hypercubes and complete graphs, but do not hold for stars or power law graphs. Additionally, Van Mieghem, Omic, and Kooij (2009) developed the N-interwined Markov Chain Model that relates the degree and the largest eigenvalue of the network and thus showed that interactions between individuals are clearly depending on the underlying structure of the network. Similar to Wang's model in Y. Wang et al. (2003), Chakrabarti et al. (2008) calculated the epidemic threshold depending on the largest eigenvalue of the adjacency matrix and came up with various policies to determine the best node to be removed from the network in order to decrease the epidemic threshold the most.

2.4 Opinion Dynamics

Another widespread application of dynamical systems on networks are opinion dynamics. The main idea behind this is that users in social networks start to adopt opinions of contacted individuals and to behave similarly to them sooner or later in time. In the real world, the processes involved in such dynamics are highly complex and depending on different factors. Since the first emergence of this idea in Weidlich (1971), many different attempts have been made to model and design these processes. The following

paragraphs describe several models based on opinion dynamics. A more comprehensive summary and in-depth few can be found in the work of Castellano, Fortunato, and Loreto (2009).

Ising Model. Named after the physicist Ernst Ising, this is a very simple and basic approach. Besides its relevance in physics, this model can also be used for opinion dynamics as stated by Binney et al. (1992). Here, agents are influenced by its interacting partners leading to order-disorder-transitions. Based on ferromagnetic interactions, the final resulting state of a system will always be ordered by one of two possible states (either positive or negative in the case of magnetism). With this in mind, the model can be applied on simple binary opinion dynamics.

Sznajd-Weron and Sznajd (2000) proposed a simple *Ising Spin Model* that can describe decision making in a closed community. Using standard Monte Carlo simulations, closely described by Mooney (1997), they found that complicated dynamics in decision making arise, finally leading to a power law in the decision time distribution. With their model they showed that in a closed community only two final states are possible: dictatorship or deadlock, meaning no common decision can be made. Furthermore, every change of an opinion leads to a further change. Whenever opinions change frequently, a period of time follows where no more decisions are made. Additionally, they described that only a small amount of the whole population can lead to a deadlock situation. However, if a group wants to win by a 50% chance, at least 70% of the whole population need to be already in consensus at the beginning of simulations.

Voter Model. This model is first explained by Clifford and Sudbury (1973) and is about the evolution and competition of different species. In Holley and Liggett (1975), it was first named and defined as the Voter Model. Its definition is fairly simple: In the initial state, each agent (or vertex in context of graph theory) is described by a binary variable and in each time step during simulations, one randomly picked agent takes the opinion of one of its randomly selected neighbors. Because of this simple definition and the urge for more realistic simulations, many modifications of the model have since been made. In the following paragraphs, some of the interesting approaches are explained.

Mobilia (2003) used an inhomogeneous voter model and introduced a *zealot*, an individual within the system that only favors one single opinion. In his work he simulated the outcome of this manipulations for 1, 2, 3 and more dimensions. He showed that in lower dimensions the introduction of the zealot results in unanimity after some time. If the dimension is equal to or greater than 3, the zealot does not effect the outcome of opinion dynamics in a noticeable way. This model was further extended in Mobilia and Georgiev (2005) and further investigated in Mobilia, Petersen, and Redner (2007), where they showed that only a few zealots introduced into the system can prevent consensus even when they face an enormous majority at the beginning of simulation.

Similar to this, Galam and Jacobs (2007) introduced *inflexibles*, again representing individuals with a fixed opinion. In their model they used normal floater agents, changing opinion based on the local majority of their neighbors, whereas inflexible agents keep their opinion throughout the whole simulation process. They observed that when using no inflexibles, the initial majority always wins by reaching consensus at the end of simulation. However, by introducing inflexibles to the simulations, an incompressible minority around the opinion where the inflexible was added starts to grow. Moreover, adding infelxibles at above a threshold of 17%, the initial minority will win.

Axelrod Model. Axelrod (1997) developed a model to simulate cultural dynamics. State transitions in this model depend on two main factors: *social influence* and *homophily*. The first one explains the phenomena of individuals becoming alike whenever they interact. The second one states that individuals that are alike tend to interact more frequently. Furthermore, nodes in a network are described by an arbitrary number of variables called *features* denoted by F, where each variable assumes an arbitrary amount of values called *traits* denoted by q. In each step during simulation, the probability of a transition, based on the overlap of a randomly selected individual and one of his neighbors, is calculated and, if high enough, one feature is adjusted to be equal. The results of simulation are strongly depending on the amount of traits as described by Castellano, Marsili, and Vespignani (2000). A small amount of traits quickly lead to consensus, whereas a large amount of traits lead to coexistence of different cultures.

There are various modifications of the original Axelrod Model. For example, Flache and M. W. Macy (2007) added a threshold that whenever the overlap of two individuals is smaller than that threshold, no adjustment of features takes place. In their work, they also pointed out that the original Axelrod model only uses nominal features, meaning that individuals are either identical or different, which can be a possible drawback.

Another approach was implemented by Centola, Gonzalez-Avella, et al. (2007), where they modified the original Axelrod Model and extended it by adding a third mechanism to the existing *homophily* and *social influence* called *network homophily*. This third mechanism allowed them to co-evolve the network structure with the cultural evolution. They found that the introduction of network dynamics does not only affect the critical value of q but also the resulting network structure. Depending on the number of possible traits, the network can evolve from a regular lattice to complex random networks or even break apart in differently sized components.

González-Avella, M. G. Cosenza, and Tucci (2005) extended the Axelrod Model to study the effects and influences of mass media on cultural evolution. Mass media influence is assumed as a fixed vector that influences the system uniformly. Given the probability B, an individual either interacts with this mass media vector as if it were a neighbor. With probability $1 - B$, the individual interacts with one of his actual neighbors. Their simulations uncovered that mass media can induce cultural diversity.

Various other modifications of the original Axelrod model showed up since its initial introduction. Klemm et al. (2003) introduced random noise to simulate *cultural drift*, the phenomena of individuals changing their opinions without external influences. Flache and M. W. Macy (2006) added *metric features*, allowing different numbers of traits q to be taken into account. A combination of a fixed vector and noise was developed and studied by Mazzitello, Candia, and Dossetti (2007). Another modification can be found in Parravano, Rivera-Ramirez, and M. Cosenza (2007), where an additional parameter limits the maximum number of shared features between individuals.

The Naming Game. Similar to the voter model and the Axelrod model, the naming game presents another way to simulate opinion dynamics. It developed from the idea drafted by Steels (1995) to explore the dynamics

and evolution of languages. Here, agents are able to develop their own vocabulary to describe, for example, physical objects but are forced to align their words whenever they are in a conversation. The simplest model based on this idea was described by Baronchelli et al. (2006). In their work, the basic rules of the naming game are defined: Every agent has an inventory of names for different objects and all inventories are empty at the beginning of simulation. At each period in time, two individuals are picked, where one is the *speaker* and the other one is the *listener*. The speaker randomly selects an object and chooses the name associated with the object of its inventory. If the inventory is empty, a new word is added to the inventory. Now the speaker transmits the selected name to the listener. If the listener has the same name in his inventory and associated to the previously selected object, the communication was successful, leaving both agents with an inventory of only that name. This means that all the other words are deleted from the inventory. However, if the listener does not have the name in his inventory, it gets appended to it and the listener associates it with the object.

There are some modified versions of this model. For example, Abrams and Strogatz (2003) tried to explain how two languages compete with each other. These two languages do not evolve over time and the more speakers on of the two languages has, the more attractive it is to individuals. In their work they found that one of the two languages always dominates, leaving the other one to be extinct sooner or later. When comparing the results of simulations to empirical data, it has shown that this model is able to predict the numbers of decreasing speakers of various endangered languages.

The model developed by Abrams and Strogatz (2003) was further extended by Minett and W. S. Wang (2008). It allows bilingualism and social structure to influence simulation results. In most cases and the absence of intervention, the dynamics of the system are equal to the original model. In addition to that, they showed that increasing the "status" of a language—meaning prestige, wealth and power of its speakers as described by Crystal (2000)—can save an endangered language from becoming extinct. However, a comparison with empirical data is missing and different aspects of the model could be refined for even better results.

More on the computational research in the field of language evolution can be found in the work of W. S. Wang and Minett (2005).

3 Materials & Methods

This Chapter gives a brief introduction to the *Activity Dynamics* framework, its variables, setup parameters and their calculation in Section 3.1, as well as a description of the used empirical data sets and the construction of collaboration networks in Section 3.2.

3.1 Activity Dynamics Framework

We use the *Activity Dynamics* framework in two ways: First, we extract user data from our empirical data sets and construct the collaboration networks used for activity simulation. Second, we simulate activity dynamics with the *Activity Dynamics* model. Here, we only cover the basic principles of the model itself. An in-depth description and a complete overview of all variables and equations can be found in Walk et al. (2015).

Based on the principles of dynamical systems on networks, the model is capable of modeling activity in online collaboration networks such as the StackExchange.com web portal. The *Activity Dynamics* model describes the activity of user i at time t as a_i. This is described in equation 3.1 and illustrated in Figure 3.1.

$$\frac{\mathrm{d}a_i}{\mathrm{d}t} = \underbrace{f_i(a_i)}_{\substack{\text{Intrinsic}\\ \text{Activity}\\ \text{Evolution of i}}} + \overbrace{\sum_j A_{ij} \underbrace{g_i(a_i, a_j)}_{\text{Influence of j on i}}}^{\text{Peer influence}}, \tag{3.1}$$

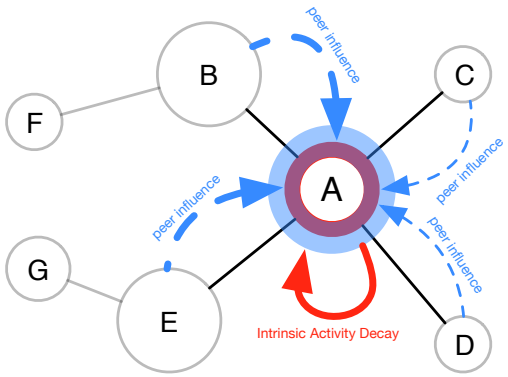


Figure 3.1: **Mechanisms of the *Activity Dynamics* Model.** This plot illustrates an example of how the mechanisms of the *Activity Dynamics* model affect one user represented by node A. The different node sizes reflect the amount of activity at one point in time. The solid lines are collaboration edges between users. This user intrinsically loses activity in each iteration (depicted by the solid red arrow and the colored, innermost circle of A), but also regains activity from neighbors through peer influence (depicted by the dashed blue arrows and the colored, outermost circle of A). If the peer influence outweighs the intrinsic activity decay, user activity increases and vice versa.

$f_i(a_i)$ specifies the intrinsic activity evolution of user i and $g_i(a_i, a_j)$ specifies the peer influence of user j on user i. Both functions are always the same for each user i and each pair of neighbors i and j. Furthermore, these two functions rely on the following two principles:

Intrinsic Activity Decay. Users tend to lose interest in certain topics while participating in online collaboration networks (for example see Danescu-Niculescu-Mizil et al. (2013)). In the *Activity Dynamics* model this is modeled as a linear function $f(a_i)$ denoted in Equation 3.2:

$$f(a_i) = -\lambda a_i, \lambda > 0, \tag{3.2}$$

where λ is the *Activity Decay Rate* specifying the decrease in activity of user i over time, provided that no external influence is exercised on this user.

Positive Peer Influence. Users copy their friends (for examples see Christakis and Fowler (2008) and Aral and Walker (2012)), that means if user j

becomes active, user i may respond to this action as well. This is covered by the sigmoid function $g(a_i, a_j)$ in Equation 3.3.

$$g(a_i, a_j) = \frac{q a_j}{\sqrt{a_c^2 + a_j^2}}, q, a_c > 0. \tag{3.3}$$

The actual amount of activity influence transferred from user j to user i depends on two parameters:

(i) The *Critical Activity Threshold* a_c describing the point at which user j has enough activity potential to notably exercise influence on user i and other neighbors of j. In systems with higher activity, a_c will be higher as well, meaning users have to be more active in order to be noticed as "more active" than the average user. In systems with low activity levels, a_c will be rather small, meaning activity of one user is recognized rather quickly. Note that peer influence on neighbors happens on all levels of activity potential.
(ii) The *Maximum Peer Activity Flow* q restricts the maximum flow of activity from user j to user i per unit time t.

The parameters, such as a_c and t, have different dimensions. Hence, Equation 3.1 is further improved by transforming it into a dimensionless form:

$$\frac{\mathrm{d}x_i}{\mathrm{d}\tau} = \underbrace{-\frac{\lambda}{\mu} x_i}_{\substack{\text{Intrinsic} \\ \text{Activity} \\ \text{Evolution of i}}} + \overbrace{\sum_j A_{ij} \frac{x_j}{\sqrt{1 + x_j^2}}}^{\text{Peer influence}}. \tag{3.4}$$

By now, the model works with only one parameter, called λ/μ. It describes the ratio between two rates:

(i) The *Activity Decay Rate* λ representing the rate of the intrinsic activity decay of user i,
(ii) and the *Peer Influence Growth Rate* μ representing the rate of activity influence user i is receiving from user j.

Table 3.1: **Model Variables and Parameters.** This table provides a quick overview of the variables and parameters used to configure the *Activity Dynamics* model.

Variable	Name	Description
λ	Activity Decay Rate	Intrinsic loss of activity a user encounters per unit time.
$\mu = \frac{q}{a_c}$	Peer Influence Growth Rate	Influence of actions taken by connected users per unit time.
a_c	Critical Activity Threshold	Soft threshold of notably influencing connected users.
q	Maximum Peer Activity Flow	Maximum activity flow from one user to another per unit time.
τ	Relative Time Scale	Represents one step in time.
Parameter	**Description**	
$\frac{\lambda}{\mu}$	The ratio between the Activity Decay Rate and the Peer Influence Growth Rate. Describes how fast one user intrinsically loses activity compared to regaining activity from neighbors.	

For example, a ratio of 10 would mean that a user intrinsically loses activity ten times faster than regaining activity from one connected user.

A short overview of all model variables and parameters can be found in Table 3.1.

Furthermore, Walk et al. (2015) conducted a *linear stability* analysis and showed that:

$$\kappa_1 < \frac{\lambda}{\mu} \tag{3.5}$$

is the master stability equation of the *Activity Dynamics* model, where κ_1 is the largest eigenvalue of the graph's adjacency matrix. Whenever the ratio λ/μ is greater than κ_1, the system will move toward the fixed point with zero activity. However, if λ/μ is smaller than κ_1, the system will become unstable and the overall activity within the network will be increased. To avoid an inactive or "dead" system, we can manipulate the system in order to violate the master stability equation denoted in Equation 3.5. We can do this in either two ways:

(i) *Manipulate the network structure* in order to change κ_1 or
(ii) *Manipulate the system's parameter* λ/μ.

Both ways of manipulation are used in our *Activity Dynamics Scenario* experiment.

3.1.1 Calculation of Model Parameters

As already described by Walk et al. (2015), the parameters for the *Activity Dynamics* model can be calculated as follows:

First of all $\mu = \frac{q}{a_c}$ is estimated, where μ is the *Peer Influence Growth Rate*, a_c the *Critical Activity Threshold* and q the *Maximum Peer Activity Flow*.

a_c is represented as the average activity per user and per period in time (in case of this work one month):

$$a_c = \frac{\sum_{t=1}^{T}(p(t) + r(t))}{nT}, \tag{3.6}$$

where T represents the totally observed time (for example 12 months), $p(t)$ represents the number of posts at time t and $r(t)$ represents the number of replies at time t.

q describes the maximum of peer induced activity that can be transferred from one user to another in one period in time. It is calculated with:

$$q = r_{\max} \frac{\sqrt{a_c^2 + \left(\frac{p_{\max}}{u_{\max}}\right)^2}}{2m\frac{p_{\max}}{u_{\max}}}, \tag{3.7}$$

where $r_{\max}$ is the maximum number of replies of all observed months, $p_{\max}$ the maximum number of posts of all observed months and $u_{\max}$ the maximum number of users of all observed months.

Finally, by linearizing around the current activity level, the ratio λ/μ can be approximated with:

$$\frac{\lambda}{\mu}(t) = \kappa_1 - \frac{1}{\mu}\log\frac{x(t+1)}{x(t)}, \tag{3.8}$$

where, again, κ_1 is the largest eigenvalue of the adjacency matrix of the graph and $x(t)$ represents the amount of activity within the collaboration network present at time t.

The estimated parameters for each empirical collaboration network can be found in Table 4.1 in Chapter 4.

3.1.2 Model Initialization

In order to simulate activity, we need to set the initial activity weights of users within the collaboration networks. Hence, we calculate the average activity per users of the first month of our observed data sets. The average is further normalized with the sum of the eigenvector centrality:

$$\overline{x} = \frac{p(0) + r(0)}{n a_c \sum_{i=0}^{n} c_i}, \tag{3.9}$$

where c_i depicts the eigenvector centrality of node i, $p(0)$ the number of posts in the first month and $r(0)$ the number of replies in the first month. We then initialize each user in the network with activity weights:

$$x_i(0) = \overline{x} c_i \tag{3.10}$$

Hence, the initial activity of a user i depends on the eigenvector centrality c_i and the overall activity $(p(0) + r(0))$ of the first month. This initialization avoids a so-called burn-in phase, where the model would require some iterations to adapt to the input values.

3.2 Empirical Data Sets

For this thesis, we have extracted a set of five different instances of the StackExchange.com networks, as well as five different Semantic MediaWiki networks. All resulting networks differ in the numbers of users and in the number of collaboration edges between them. Table 3.2 lists a detailed overview of the exact numbers for each of the extracted networks.

We use these empirical data sets for the simulation of activity, to study the impacts of our six different plausible real-world scenarios on overall network

Table 3.2: **Characteristics of Empirical Data Sets.** The five StackExchange.com networks and five Semantic MediaWikis all differ in the number of users, the number of collaboration edges and the number of interactions (posts and replies). We simulate activity over the last twelve months of each data set (as stated by *Start* and *End* columns) for all three experiments conducted as part of this work. In the case of the *Activity Dynamics Scenarios* experiment, we simulate activity with affected users at the beginning of month eight.

Dataset	**BeerStack-Exchange**	**BitcoinStack-Exchange**	**ElectronicsStack-Exchange**	**PhysicsStack-Exchange**	**GamingStack-Exchange**
Users	469	6,314	22,064	23,834	34,701
Edges	1,198	17,842	121,205	129,615	132,414
Posts	199	2,755	16,484	21,217	13,122
Replies	1,190	12,740	121,573	136,190	65,642
κ_1	19.651	56.667	154.153	145.883	141.345
Avg. Degree	5	6	11	11	8
Start	Feb 2014	Feb 2014	Feb 2014	Feb 2014	Feb 2014
End	Feb 2015	Feb 2015	Feb 2015	Feb 2015	Feb 2015

Dataset	**ComplexOperations**	**BioInformatics**	**NeuroLex**	**DotaWiki**	**PracticalPlants**
Users	285	308	1,183	2,023	2,220
Edges	452	314	1,875	4,048	148
Posts	181	207	11,567	244	2,330
Replies	3,896	135	25,061	2,329	15,481
κ_1	10.680	10.549	27.415	35.298	10.626
Avg. Degree	3	2	3	4	1
Start	April 2013	March 2013	Nov 2012	April 2012	Sep 2012
End	April 2014	March 2014	Nov 2013	April 2013	Sep 2013

activity and to conduct the correlation analysis between centrality measures of users and simulated activity. Furthermore, we evaluate the simulation performance of the introduced dynamic network structure by comparing it with the original static network structure of the *Activity Dynamics* model. A detailed description of the three conducted experiments can be found in Chapter 4 to Chapter 6.

For all simulations of our three experiments, we consider the last 12 months of the corresponding data set. For our *Activity Dynamics Scenarios* experiment, simulations of each scenario approach starts at the beginning of month 8. Hence, the first results of simulation for all scenarios start at month 9.

Posts for the five StackExchange.com networks are defined as asking ques-

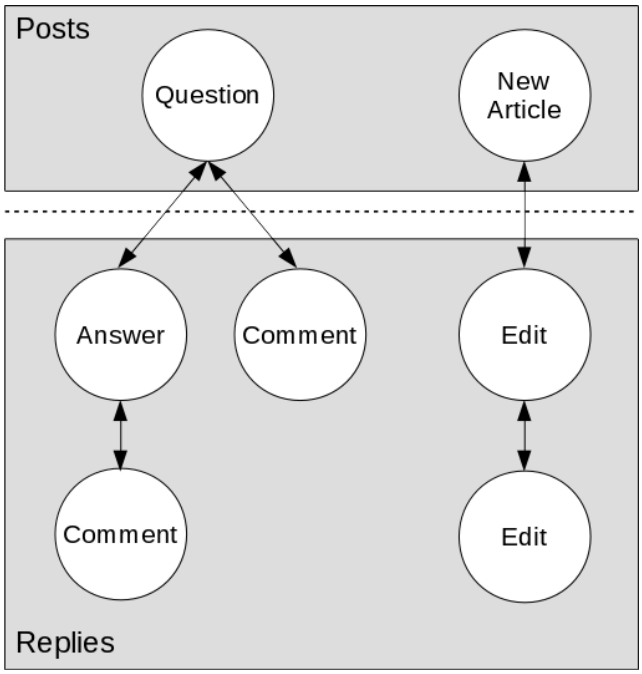
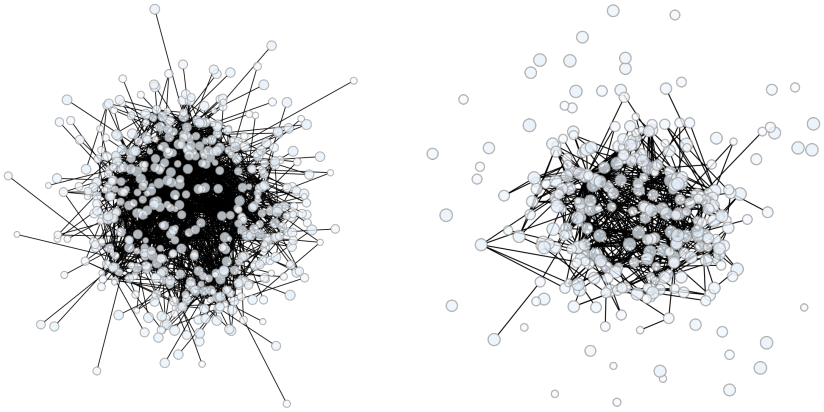


Figure 3.2: **Construction of Collaboration Networks.** This figure depicts how posts and replies are defined for all StackExchange.com networks (left side), all Semantic MediaWikis (right side) and how collaboration networks are constructed based on the empirical data sets listed in Table 3.2.

tions, while *replies* consist of answers and comments. In the case of the five Semantic MediaWiki instances, this definition is different. Here, the creation of an article counts as a post, while the edit to an existing article counts as a reply.

Collaboration for the StackExchange.com data sets is defined as users either posting an answer to a question, or commenting on an answer or a question. For the Semantic MediaWiki data sets, collaboration between two users is given when they subsequently worked on the same article. For example, if user i creates an article that is then edited by user j or user i edits an article after user j has edited that same article, a collaborative edge between these users is created. These mechanics are described in Figure 3.2.

In general, these networks can be represented as an undirected graph $G = \{V, E\}$ with a set of nodes (users) V and edges (collaboration ties) E



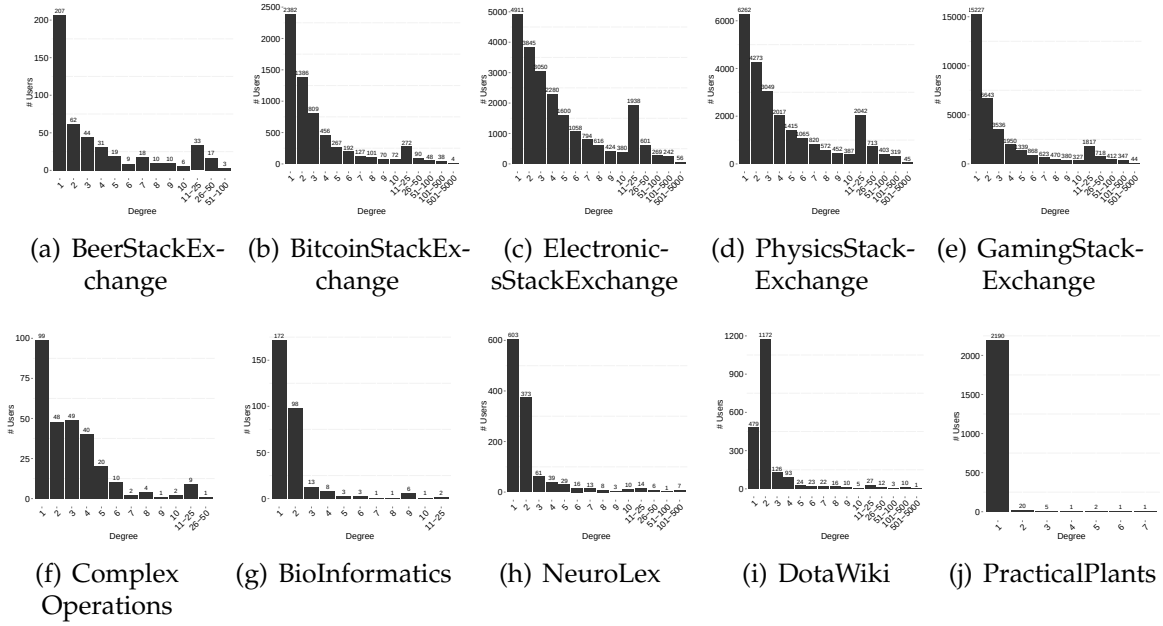
(a) BeerStackExchange (b) ComplexOperations

Figure 3.3: **Exemplary Resulting Collaboration Networks.** Representation of the resulting collaboration networks in the form of graphs for the *BeerStackExchange* network and the *ComplexOperations* network. Users are represented as nodes and collaborations between them as edges. Node sizes represent the initial empirical activity at the beginning of simulation. *ComplexOperations* (b) illustrates unconnected nodes, meaning they have not collaborated during the observed time span.

between these nodes. One way to represent such graphs algebraically is as an $n \times n$ *adjacency matrix* $\boldsymbol{A}$, with n being the total number of nodes in the network. If two users i and j have collaborated in the past, then $A_{ij} = 1$, connecting them via an (collaboration) edge, and $A_{ij} = 0$ otherwise so i and j are not connected in the resulting graph. Note that collaboration networks are undirected, making $\boldsymbol{A}$ symmetric with the total number of links m for the collaboration networks being defined as $m = \frac{1}{2} \sum_{ij} A_{ij}$.

Figure 3.3 depicts exemplary resulting collaboration networks for one of our StackExchange.com networks and one of our Semantic MediaWiki data sets in the form of graphs. These graphs consist of nodes that represent users and edges that represent collaborations between them. Based on the empirical data, the graphs differ in the amount of nodes and edges.

Furthermore, Figure 3.4 depicts the degree distributions (numbers of collaborations with other users per user) among users for all our StackExchange.com data sets and for all our Semantic MediaWiki data sets. The amount of collaborations among users differs for all networks. The major



(a) BeerStackExchange (b) BitcoinStackExchange (c) ElectronicsStackExchange (d) PhysicsStackExchange (e) GamingStackExchange

(f) Complex Operations (g) BioInformatics (h) NeuroLex (i) DotaWiki (j) PracticalPlants

Figure 3.4: **Degree Distributions of Collaboration Networks.** This Figure illustrates the distribution of the users' degrees for all our extracted StackExchange.com data sets (top) and our Semantic MediaWiki data sets (bottom). Numbers above bars represent the amount of users in each bin. All networks contain a large amount of users with only a few connections to other users, whereas only a rather small amount of users collaborated with many other participants.

part of users of all our data sets have only collaborated once or twice with other users during the observed period of time. For example, in the *BeerStackExchange* network (Figure 3.4(a)), almost half of all existing users only interacted with one other unique user. In the *ComplexOperations* network (Figure 3.4(f)), almost all users showed similar behavior and only collaborated with one other user. Each extracted network contains only a few users that collaborated with many other unique users and are therefore more important within the graph.

StackExchange.com data sets. StackExchange.com provides its users a question and answer portal for a variety of topics. Participating users seek answers to their posted questions and can further discuss their issues by commenting on answers and questions. Additionally, the relevance of posts

can be marked by up or down votes. Each available topic is treated in a discrete subnetwork. We decided to use five different instances of these subnetworks to cover diverse forms of collaboration networks. The *Beer-StackExchange* network is a portal for beer lovers and people interested in the process of brewing beer. This data set represents the smallest one of all used StackExchange.com networks with only 469 active users and 199 posts and 1,190 replies at the end of our record. The *BitcoinStackExchange* network is about topics related to the online payment system and virtual currency Bitcoin. With 6,314 active users and 2,755 posts and 12,740 replies, this data set is one of our smaller empirical networks we used in this thesis. Slightly larger and a bit more active is the *ElectronicsStackExchange* network with its 22,064 users, creating a total amount of 138,057 overall activity (posts + replies) about the fields of Electrical Engineering. Almost identical in the number of users is the *PhysicsStackExchange* network, where a total of 23,834 users, mainly researchers of Physics, engaged in 21,217 posts and 136,190 replies. The largest network among our five StackExchange.com networks is the *GamingStackExchange* network, also known as Arqade, where 34,701 users interested in topics related to video games contributed 13,122 posts and 65,642 replies.

Semantic MediaWiki data sets. Semantic MediaWikis—often referred to as "Wikis"—are online encyclopedias about arbitrary topics. Their visitors can browse through articles and gain information about the topics they are interested in. Everyone is allowed to participate and create or edit articles in almost any of those networks. Again, we cover different sizes of Semtantic MediaWiki networks. On *ComplexOperations*, 285 users created 181 articles that have been edited 3,896 times. It is mainly about political conflicts all around the world and is the smallest empirical data set we have used. Another rather small network is the *BioInformatics* Wiki with a total amount of 308 users, that have created 207 articles related to the field of Bioinformatics. These articles have since been edited 135 times. The *NeuroLex* network is a medium-sized data set and provides a lexicon about Neuroscience. 1,183 active users engaged in an overall activity of 36,628, with 11,567 created articles that have been edited 25,061 times. On *DotaWiki*, German gamers of the multiplayer online battle arena game Dota find various tips in how to improve their gameplay and master tough challenges. With 2,023 active users that have contributed to the network, it

is among our larger Semantic MediaWiki data sets. 244 articles have been created and edited 2,329 times. The largest network examined in this thesis is the *PractlicalPlants* Wiki. It offers information about useful plants in 2,330 articles that have been edited 15,481 times, contributed by a total of 2,220 users.

4 Activity Dynamics Scenarios

In this chapter we present our six different plausible real-world scenarios. We particularly study and focus on these six scenarios due to their consequential implications for activity dynamics in online collaboration networks. The idea behind these scenarios is to sketch plausible real-world events that could take place on collaborative websites, such as the StackExchange.com network, and to ultimately provide guidance for website owners based on the results of our simulations.

Two of the scenarios deal with rapid decrease and increase in user numbers. Further two scenarios cover cases in which users suddenly start or stop collaborating with each other. Additionally, we conduct one scenario which covers situations in networks, in which users do not lose interest in participating. Such users can, for example, be moderators paid by website owners trying to abet other users to engage in activity. Furthermore, we simulate the occurrence of trolls—individuals trying to irritate the network's users—in our empirical networks. The following sections describe all six scenarios in detail and give examples of real-world incidents.

The simulation of each scenario starts at month 8 for all empirical data sets. Hence, the first results of manipulated activity dynamics can be seen at the beginning of month 9. Additionally, we simulate activity with non-manipulated networks from month 1 to month 8 in order to demonstrate the accurate simulations of the *Activity Dynamics* framework.

Table 4.1 lists all parameters needed for simulation, calculated as described in Section 3.1.1, respectively for all StackExchange.com data sets and all Semantic MediaWiki data sets.

We select affected users and collaboration edges between these users for all six scenarios in two different ways. This leaves us with two different approaches for each scenario:

Table 4.1: **Model Parameters for Each Empirical Data Set.** This table lists all parameters resulting from the calculations, described in Section 3.1.1, for each of our StackExchange.com networks and our Semantic MediaWiki networks, listed in Table 3.2. For all data sets, the time spans have been set to $t = 1$ month and $\Delta\tau = 0.001$. Additionally, the ratio λ/μ is only listed for the first month of simulation.

Dataset	BeerStack-Exchange	BitcoinStack-Exchange	ElectronicsStack-Exchange	PhysicsStack-Exchange	GamingStack-Exchange
a_c	0.228	0.189	0.481	0.508	0.175
q	1.341	1.013	1.317	1.404	0.692
p_{max}	0.016	0.028	0.03	0.032	0.023
$\Delta\tau$	0.001	0.001	0.001	0.001	0.001
μ	5.887	5.367	2.736	2.764	3.965
$\frac{\lambda}{\mu}_{init}$	19.802	56.675	154.073	145.843	141.306

Dataset	ComplexOperations	BioInformatics	NeuroLex	DotaWiki	PracticalPlants
a_c	1.1	0.085	2.382	0.098	0.617
q	11.259	5.199	39.263	0.838	142.063
p_{max}	0.454	0.343	3.397	0.054	0.133
$\Delta\tau$	0.001	0.001	0.001	0.001	0.001
μ	10.232	60.865	16.486	8.561	230.191
$\frac{\lambda}{\mu}_{init}$	11.062	10.518	27.476	35.5	10.631

(i) *Randomly* picking users or collaboration edges and presenting the average results over a total of 10 random iterations or
(ii) Performing *informed* selections of users and collaboration edges based on a preference towards the highest degree of the corresponding users.

Furthermore, if the structure of the collaboration network is changed by the implementation of a scenario, we update κ_1 and subsequently the ratio λ/μ for the simulation of months 8 to 12, assuming that the numbers of posts and replies remain unaffected.

4.1 Mass Emigration

Nowadays there are countless providers of social networking platforms (for example Facebook, Twitter, Instagram, Youtube or Tiwtch) and due to this huge amount of accessible services, such websites often struggle to keep their users and not to lose them to competitors. However, due to different reasons, such as the natural growing desire for “something new”

or a change in the terms-of-service or privacy agreements upsetting users, a sudden decrease of user numbers might occur. This might be because users have lost interest in participating in such networks or just because they changed to another social network.

theguardian.com (2013) wrote that the online social networking site Facebook lost millions of users per month due to upcoming competitors and alternative social networks such as Twitter. Even though its one billion active users, drops in activity are noticeable in some places around the world. Another real-world example is described by wired.com (2013), where Instagram—an online service for sharing photos with friends—recorded massive drops in user numbers after they updated their terms-of-service and parts of their privacy policy. New regulations allowed the company to sell users' photos and information to advertising companies. It ultimately led to the loss of daily active users and some users even quit Instagram altogether.

Implementation. For this specific scenario, we assumed that one such event triggers the desire of multiple users to seize contributing to a specific collaboration network. For the *random* approach a total of 1%, 5% and 10% of all existing users are selected and entirely removed from the graph. Analogously, for the *informed* approach, we select users with the highest number of collaboration edges—specified by a user's degree—for removal. Furthermore, all collaboration edges to and from these removed users are removed as well. Hence, the resulting manipulated network has less users that could engage in collaborations and less collaboration edges between them so peer influence is limited additionally.

4.2 Mass Immigration

Newly added features or pre-planned events, such as competitions, sales promotions or maybe even the release of a new version of a collaborative website, force administrators and website owners to cope with a sudden increase of user activity on their websites caused not only by existing users but also by newly joined community members. This sudden increase in user activity could lead to overloaded servers and lags for visitors that

dampen the user's experience. On supervised networks, administrators could drag behind new posts and overlook forbidden content that could imply legitimate problems or forfeit in overall quality of content.

theverge.com (2013a) described a similar incident that happend to the Canadian telecommunication company BlackBerry Limited in the year 2013. When the company was waiting in the wings for the release of their instant messaging app BBM (BlackBerry Messenger) to iOS Appstore and Google Play, a leaked version could illegally be downloaded over the Internet before the official release. This version was downloaded over one million times and BlackBerry was not able to cope with the sudden rush of newly joined users. Hence, all servers were overloaded, eventually causing the whole service to crash.

Implementation. We introduce a total of 1%, 5% and 10% from the total amount of nodes present in the corresponding collaboration network as new users an connect them with existing ones for this particular scenario. In the *informed* approach, we connect these new users specifically to existing users with the highest number of collaboration edges (highest degree), while in the *random* approach we do not select specific target users and new users are just connected randomly with existing ones.

The number of collaborative edges for each newly added user is equal to the floored average degree of the corresponding collaboration network (See Table 3.2 in Section 3.2).

4.3 Breaking Collaborative Ties

This scenario is similar to the *Mass Emigration* scenario. Users might exhibit a change of interests during their membership in a collaborative network. However, instead of completely leaving the network, they just stop collaborating with a selection of their currently connected users and only stay in contact with users that participate in the same field of interest. Another possible trigger for such user behavior could be dispute and disunity among connected users, as well as harassment.

The huffingtonpost.com (2011) tried to find out the answer to why users unfriend on Facebook and hence break their collaborative ties. According to this article, offensive comments and lack of knowledge about a person are the top reasons why users unfriend. Other reasons are depressive comments or comments with political contents. So it appears that users easily connect with other users they already know through real-life, but start to break these connections if interests between them turn out to be too diverse over time.

Implementation. We reproduce this occurrence by removing a total of 10%, 30% and 50% of all existing collaboration edges between existing users. Again, we randomly select edges for removal in the *random* approach, whereas edges of users with the highest amount of connections to other users are selected for removal in the *informed* approach. The selected edges are removed entirely from the graph, which causes a decrease of κ_1.

4.4 Establishing New Collaborations

Whenever users decide to participate in an online collaboration network, they interact with existing users and create collaborative links to them. While these links naturally emerge and evolve over time, for example, if users share the same field of interest or know each other already outside of the network, these connections could also be actively promoted. Events like workshops that concentrate on a specific topic could bring together more people in a shorter time frame. Moreover, new connections could be created if certain users are featured and introduced to other users through the implementation of recommendation systems. Such systems find users with similar interests, based on a variety of parameters, and implement ways to promote interactions between a network's users.

An example of a user recommendation system implemented by Facebook is given by washingtonpost.com (2015), where they explain how Facebook suggests people one might know in real-life but are not yet connected in the network. Whenever a user connects to another user, Facbook uses the emerging network structure and calculates statistical possibilities for other users you are most likely to know and connect to in the future. The individually predicted users are then shown to each existing user under

"People You May Know". Hence, Facebook accelerates the creation of new collaboration edges between existing users, eventually increasing activity.

Implementation. In this scenario, we assume that an administrator introduces such a user recommendation system to our empirical networks. This system suggests users to get in contact with other users based on shared fields of interests. In particular, we create and add a total of 10%, 30% and 50% of all existing collaboration edges to the collaboration network. First, we *randomly* select a source and a target user and add a new collaboration edge between them if they are not yet connected. Second, we select the users with the highest amount of collaboration edges and connect them with each other until the required amount of newly created edges is reached.

4.5 Providing Incentives

Websites that struggle with low levels of overall activity might introduce (monetary) incentives for existing users or external experts that provide a constant output of activity to the network. For example, these experts might be employed to answer questions on one of the StackExchange.com networks or to write high quality articles on one of the Semantic MediaWiki networks. Additionally, they could perform administrative tasks, such as moderation of parts of the website. However, this scenario does not aim to simply create new content for users, but also to create an active environment where users are motivated to participate and engage in collaborations with other participants. In an ideal case, the overall activity of the network becomes self-sustainable and the initially added incentives are not needed anymore.

techcrunch.com (2007) and theverge.com (2012) give a real-world example of what incentives in online social networks could look like. They described the introduction of YouTube's *Revenue Sharing Partners Program* that allows creators of popular videos to partner up with YouTube and to monetize their content. These users gain special privileges such as in-video advertising or special options in designing their channels. This program aims to motivate already popular creators to upload even more content, as both YouTube and

the creators will benefit from the higher amount of advertising they can sell to their viewers.

Implementation. For simulating monetary incentives, we select a total of 1, 5 and 10 targets among all existing users and increase their activity every month by $\frac{10}{na_c}$. This means that an additional 10 posts or replies per month per incentivized user are introduced to the network to increase overall activity. Again, we first select affected users *randomly* among all existing users and then specifically select the users with the highest amount of collaboration edges for the *informed* approach.

Furthermore, to avoid a sudden increase of activity at the beginning of each month, we equally distribute the additionally introduced activity of each incentivized user over the whole time interval t.

4.6 Emergence of Trolls

Websites with any kind of user generated content are often exposed to users that want to harm the network by engaging in activities that upset other participants. Such users—referred to as trolls— typically try to interrupt and disturb discussions and steer them off-topic by creating spam or questionable and highly controversial topics with only one aim: To lure other users into meaningless discussions. For example, in one of our StackExchange.com networks, trolls might intentionally give wrong answers to set users on the wrong track. In Semantic MediaWikis, trolls might falsify articles on purpose so that other users need to react to keep the desired quality of content. Whenever such trolls emerge in a collaboration network, we assume overall activity to raise due to the reactions of other users. However, we argue that the activity resulting from these reactions is not the kind of activity that website owners and administrators want to see in their networks. In fact, we assume that users answering trolls waste their time and energy where they normally could have created meaningful contributions.

theverge.com (2013b) stated that trolls are affecting they way how other users feel about a subject of an article. Additionally, they described that the “less civil” those troll comments are, the bigger will be the impact on

thoughts of other users. These impacts can reinforce opinions that users already have. Furthermore, theverge.com (2015) depicted how Twitter is still having problems in dealing with trolls. Twitter CEO Dick Costolo said that posts of trolls cost core users that provide actual content, as they leave the network frustrated. As long as they do not find beneficial ways to ban trolls, meaningful activity will decrease over time.

Implementation. To represent this unwanted activity and the process of "wasting" time on trolls, we add a total of 1, 5 and 10 users to the collaboration networks and set their activity to $-\frac{10}{na_c}$. This is equivalent to ten unwanted posts (expressed by the negative activity weight).

Similar to the previously mentioned scenarios, we at first *randomly* connect newly introduced trolls to existing users and then specifically connect them to users with the highest amount of collaboration edges for our *informed* approach. We set the amount of connections between introduced trolls and existing users to the floored average degree of all existing users in the network (See Table 3.2 in Section 3.2). As administrators usually quickly deal with trolls in supervised networks, we only initialize the trolls once with negative activity. Additionally, we suspend the *Activity Decay Rate* for these nodes, as the negative influence will turn positive otherwise (minus times minus would result in a plus). Thus, the introduced trolls exercise a negative influence on their neighbors who then waste their time and energy on the troll. As long as the negative activity of a troll has not reached 0, peers of the troll positively influence his activity. When the activity reaches a level of 0, the activity of a troll is no longer changed for the rest of simulation.

5 Centrality Analysis

According to the way we create collaboration networks, an important factor of user activity in such networks is the underlying network structure. More active users potentially interacted with a larger amount of other users in the network and are therefore more central than users with less activity. This fact has a major impact on the peer influence received from neighbors. We investigate the underlying network structure in combination with the *Activity Dynamics* framework with the intention to find valid statements about the evolution of activity that can be made by only considering a network's underlying structure. Hence, the need of activity dynamics simulations might become unnecessary.

The construction of collaboration networks in the *Activity Dynamics* framework is influencing the resulting amount of collaboration edges and thus the centrality of users. Hence, we manipulate this construction to evaluate if different types of collaboration networks correlate unequally with simulated activity. The following sections describe the two processes included in this experiment. First, we create three different types of collaboration networks based on the number of interactions between users for each of our data sets. Second, we calculate correlation coefficients based on three different centrality measures and simulate activity for each of the collaboration networks resulting of the first step.

5.1 Construction of Collaboration Networks

As mentioned in Section 3.2, we create the collaboration networks based on the empirical data we used in this thesis. Whenever two users interacted at least once during the observed period of time, we connect these two users

Table 5.1: **Changes in the Amount of Collaboration Edges.** This table depicts the changes in the number of edges of the resulting graphs depending on k. The percentages in braces list how much edges are left compared to $k = 1$, where 100% of possible edges exist. The changes are listed for all empirical data sets we used in this thesis. The number of edges is decreasing with higher values of k, proving that only a few users collaborated more than once with each other.

Data Set	$k = 1$	$k = 2$	$k = 3$
BeerStackExchange	1,198	216 (18.03%)	85 (7.1%)
BitcoinStackExchange	17,842	5,434 (30.46%)	2,377 (13.32%)
ElectronicsStackExchange	121,205	45,827 (37.81%)	22,366 (18.45%)
PhysicsStackExchange	129,615	45,579 (35.16%)	21,822 (16.84%)
GamingStackExchange	132,414	37,976 (28.68%)	16,046 (12.12%)
ComplexOperations	452	252 (55.75%)	159 (35.18%)
BioInformatics	314	100 (31.85%)	48 (15.29%)
NeuroLex	1,875	727 (38.77%)	456 (24.32%)
DotaWiki	4,048	1,039 (25.67%)	527 (13.02%)
PracticalPlants	148	60 (40.54%)	33 (22.3%)

with an undirected collaboration edge. For this experiment, we manipulate the number of required interactions between two users in order to link them together in the first step of this experiment. The presupposed number of needed interactions is denoted by k. With this in mind, we set $A_{ij} = 1$ only if the number of collaborations between two users is equal to or greater than k. Hence, the resulting minimum degree of users within the graph will always be equal to k. This ultimately results in a collaboration network with a few centralized users and many unconnected users.

In this first step, we reconstruct each collaboration network for each empirical data set where $k = 1$, $k = 2$ and $k = 3$. We did not consider higher values for k, as the resulting graphs would have an insufficient amount of collaboration edges. Note that the number of nodes stays unaffected and only the resulting amount of collaboration edges changes with k (see Table 5.1). Furthermore, $k = 1$ is equivalent to the original process of constructing the collaboration networks in the *Activity Dynamics* framework, hence the numbers of edges are the same as in Table 3.2 in Section 3.2.

5.2 Correlation with Centrality Measures

In order to make general statements of how the centrality of users is affecting the simulation of activity performed by the *Activity Dynamics* framework, we conduct a correlation analysis based on three different centrality measures. First, we investigate the correlation between simulated activity and the degree of nodes. Second, we calculate and illustrate the correlation of simulated activity and eigenvector centrality of nodes. Third, we show relations between the simulated activity and Google's PageRank. We explain the differences between these three centrality measures in the following paragraphs.

To calculate the correlation coefficients, we calculate the sum of activity weights per user over the whole process of simulation (month $1 - 12$) and then compare this sum with each of the three centrality measures.

Degree Centrality. As explained by Gross and Yellen (2005), the degree of a node x of a graph is denoted by $deg(x)$ and represents the number of edges connecting a node to its neighbors plus twice the number of self-loops if present. This is the simplest centrality measure for nodes in a graph.

Eigenvector Centrality. More significant centrality values can be calculated through the eigenvector centrality. Unlike the degree centrality, where every connection is equally weighted, the eigenvector centrality weights connections based on their centralities. Hence, the whole structure of the network is taken into account, which is elaborately explained by Bonacich (2007). Equation 5.1 specified in the work of Newman (2010) describes the eigenvector centrality of node x as:

$$x_i = \kappa_1^{-1} \sum_j A_{ij} x_j, \tag{5.1}$$

where κ_1 is the largest eigenvalue and A the adjacency matrix of a graph.

PageRank. Invented by Google's founders Larry Page and Sergey Brin, the PageRank algorithm is another way of calculating the importance of nodes within a graph. It was initially designed to measure the importance of websites on the Internet by taking various factors into account in order

to better fit people's subjective idea of importance, it can be also used on the resulting graphs we construct as part of this work.

The PageRank of website A is defined in Brin and Page (2012) and depicted in Equation 5.2.

$$PR(A) = (1-d) + d(PR(T_1)/C(T_1) + ... + (PR(T_n)/C(T_n)), \tag{5.2}$$

where parameter d is a damping factor ranging from 0 to 1 and usually set to 0.85, as larger values of d would require more iterations during calculations. $T_1...T_n$ are all websites pointing to A and $C(A)$ is the number of pages website A points to. This algorithm can be calculated by using an iterative approach and forms a probability distribution over all websites, hence the sum of the PageRank of all websites will always be 1.

In context of graph theory, websites are represented by nodes and links between websites are represented as directed edges. In case of this work, where we only use undirected graphs, the PageRank is equal to the degree distribution of the graph as described by Newman (2010).

Pearson Correlation Coefficient. In order to measure correlations between the aforementioned centrality measures and simulated activities, we calculate the Pearson Correlation Coefficient. This coefficient was developed by Karl Pearson and describes the linear relationship between two quantitative variables X and Y. Equation 5.3 shows its definition as described by Kirk (2007):

$$\rho = \frac{\frac{\sum(X_i-\overline{X})(Y_i-\overline{Y})}{n}}{\sqrt{[\frac{\sum(X_i-\overline{X})^2}{n}][\frac{\sum(Y_i-\overline{Y})^2}{n}]}}. \tag{5.3}$$

The resulting value of ρ ranges between -1 and $+1$, and states two facts about the relationship between X and Y:

(i) Its *strength*, that represents how strong the relationship between the two variables is and
(ii) its *direction*, stating if the relationship is positive ore negative, represented by the sign of ρ.

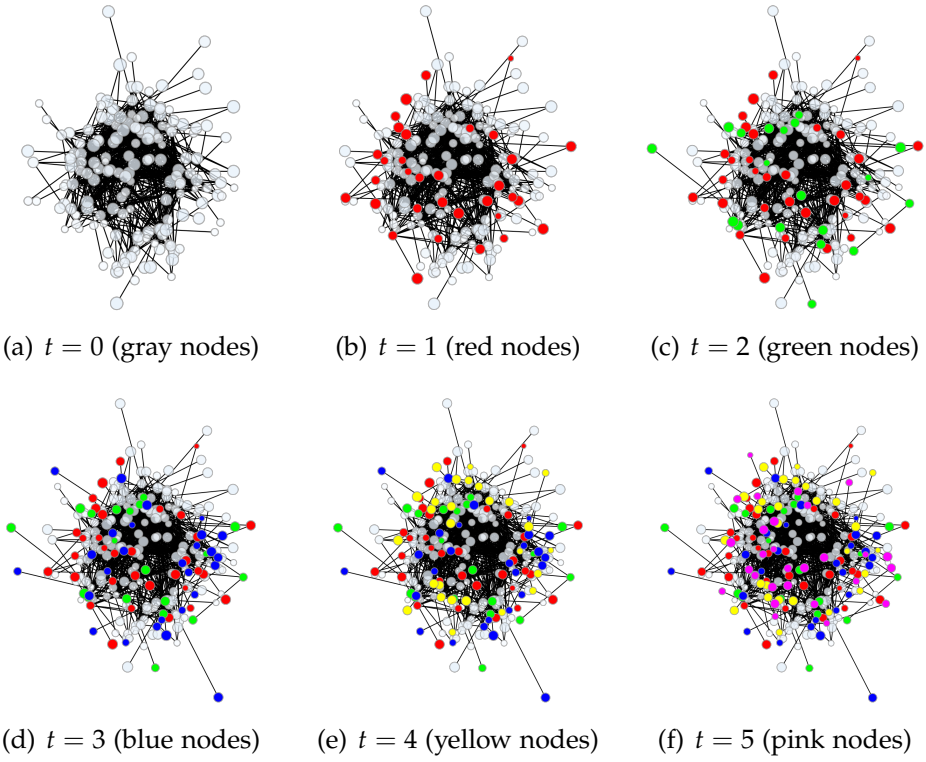
This means that if $\rho = 1$ a total positive correlation is given, when $\rho = -1$ a total negative correlation is given and when $\rho = 0$ no correlation at all is given.

6 Dynamic Network Structure

In the original *Activity Dynamics* framework, a static collaboration network based on an empirical data set is constructed and used for the calculation of parameters and for simulation of activity dynamics. The parameters integrated in the calculation of the ratio λ/μ are only calculated once at the beginning of the simulations and do not adopt changes in the underlying network structure over time. In addition to that, all users within the collaboration network are already part of the graph at the beginning of simulations ($t = 0$) even though they might have joined the network at a later point in time (for example $t = 4$). However, as collaboration networks are highly dynamic and evolve over time, more users join the network, existing ones connect with other users and therefore the underlying structure of the graph is ever changing over time. We assume that this approach is subject for improvements in order to gain an even more accurate simulation performance.

For this third and last experiment, we introduce a dynamic network structure that adopts changes in user numbers and in the amount of collaboration edges over time. We use this dynamic structure to calculate model parameters and to simulate activity. Hence, the underlying structure of the collaboration network changes per period in time t, initially starting with a graph that only contains users that actually interacted in the first month ($t = 0$). As time evolves, more and more nodes and collaboration edges are added to the graph. Furthermore, we calculate all parameters that are included in the calculation of the ratio λ/μ per period in time t.

The following sections describe the changes in data preprocessing, differences in calculation of the model parameters and variations in the initialization of the *Activity Dynamics* model.



(a) $t = 0$ (gray nodes) (b) $t = 1$ (red nodes) (c) $t = 2$ (green nodes)

(d) $t = 3$ (blue nodes) (e) $t = 4$ (yellow nodes) (f) $t = 5$ (pink nodes)

Figure 6.1: **Evolution of the BeerStackExchange Network Structure.** Visualization of the dynamic network structure of the *BeerStackExchange* data set for the first 6 months ($t = 0...t = 5$) of simulation. Users and edges are added to the graph as time evolves, with colors of the nodes representing the month they were added.

6.1 Changes in Data Preprocessing

The *Activity Dynamics* framework is already capable of preprocessing empirical data sets and generating graphs that are then used for the simulation of activity dynamics. However, we made small changes to that preprocessing in order to take a dynamic network structure into account. Our used data sets contain time stamps of each interaction between users. These time stamps come in handy to examine the changes in user numbers and the amount of links between them that take place over time. Therefore, we manipulate the construction of graphs to add these time stamps as node properties that

can further be extracted during simulations. The model uses the extracted time stamps to filter nodes and collaboration edges automatically for each of the observed 12 months during the simulation process. Hence, the underlying structure of each network is different for each month. We applied the manipulated preprocessing to all five StackExchange.com networks and all five Semantic MediaWikis. Figure 6.1 illustrates an example of the monthly evolution of the *BeerStackExchange* network over the first 6 months.

6.2 Changes in Parameter Calculation

With the changes made to the underlying network structure, we also have to recalculate all variables and parameters (described in Section 3.1.1) for each of the 12 months. Originally, we calculated these parameters by either using averages or summed up values of the observed time span in the original "static" approach. This is different with the dynamic network structure, as we now have values, such as the number of users or the number of posts and replies, that differ for each month. Hence, the calculation of these parameters changes compared to the original implementation of the *Activity Dynamics* framework and resulting parameters are different for each month, always depending on time t. We cover all changes in calculations in the following paragraphs.

The single parameter—the *ratio* λ/μ needed for model setup—is calculated for each month as in the original approach. However, we now estimate $\mu = \frac{a_c}{q}$ for every unit time t transforming it to $\mu(t) = \frac{a_c(t)}{q(t)}$, leaving us with a different μ for each month.

The *Critical Activity Threshold* a_c does no longer sum up the number of posts and replies over the whole time span. It now covers the actual numbers of posts and replies of each month, thus we changed it to:

$$a_c(t) = \frac{(p(t) + r(t))}{u(t)}, \tag{6.1}$$

where $p(t)$ and $r(t)$ are the number of posts and replies at time t. $u(t)$ is the number of users present in the network at time t.

The *Maximum Peer Activity Flow q* has taken the maximum number of posts, replies and users per month and the whole time span into account. However, we can now change this to consider the actual number of users, edges and users per month. Hence, the calculation of q changes to:

$$q(t) = r(t)\frac{\sqrt{a_c(t)^2 + \left(\frac{p(t)}{u(t)}\right)^2}}{2m\frac{p(t)}{u(t)}}, \tag{6.2}$$

where, again, $r(t)$ is the number of replies at time t, $p(t)$ is the number of posts at time t and $u(t)$ is the number of users present in the network at time t.

With these updated variables, the ratio gets approximated in the same way as described in Section 3.1.1, but with a changing μ and κ_1 for each month ($t = 0$ to $t = 11$):

$$\frac{\lambda}{\mu}(t) = \kappa_1(t) - \frac{1}{\mu(t)} \log \frac{x(t+1)}{x(t)}. \tag{6.3}$$

Note that $\kappa_1(t)$—the largest eigenvalue of the graph's adjacency matrix—needs to be updated at every period in time t due to the dynamic changes in the underlying network structure performed by adding nodes and edges to the graph, always depending on empirical data. $x(t)$ represents the amount of activity within the collaboration network at time t.

6.3 Changes in Model Initialization

We also change the way of how activity weights of nodes are initialized compared to the original approach described in Section 3.1.2. Originally, all users contained in the empirical data set are represented as nodes in the graph right from the beginning of simulation, even though they might occur at a later point in time. Thus, all activity weights of users are only initialized once, where the activity weight of each user depends

on the empirical activity they actually exercised in the first month ($t = 0$) of the complete observed time span. With the possibility of using a dynamic network structure, we adapt this approach and initialize nodes at the beginning of each month.

As in the original "static" approach, we calculate the average activity of all nodes that are present in the graph, but this time for each month. This average activity results from the empirical activity of newly joined users (i.e. users that have not been present in the graph before) and the simulated activity of previously existing nodes if it is not the first month of simulation ($t > 0$). So Equation 3.9 depicted in Section 3.1.2 changes to:

$$\overline{x}(t) = \begin{cases} \frac{p_n(t)+r_n(t)}{na_c(t)\sum_{i=0}^{n} c_i(t)}, & \text{if } t = 0 \\ \\ \frac{p_n(t)+r_n(t)+x_{sim}(t-1)}{na_c(t)\sum_{i=0}^{n} c_i(t)}, & \text{if } t > 0 \end{cases} \tag{6.4}$$

where $c_i(t)$ is the eigenvector centrality of node i at time t and $p_n(t)$ and $r_n(t)$ are the numbers of posts and replies of *newly added* users at time t and $x_{sim}(t-1)$ is the sum of simulated activity of all users already present in the previous month (for $t > 0$). Then we can (re-) initialize each node i in the network with activity weights x at each point in time t with:

$$x_i(t) = \overline{x}(t)c_i(t). \tag{6.5}$$

Note that due to the fact that simulated activity x_{sim} is depending on the number of users present in the graph at time t and the *Critical Activity Threshold* $a_c(t)$ at time t, we have to update x_{sim} between the simulations of each month. Hence, we update x_{sim} every time the simulation of one month is finished and the simulation of the next month begins. We do this by multiplying the simulated activity of t with the current $a_c(t)$ and by dividing it with the new $a_c(t+1)$ of the next month $(t+1)$ and by further multiplying it with the current number of users of t and by then dividing it with the number of users present in the next month $(t+1)$ as denoted by Equation 6.6:

$$x_{sim}(t) = x_{sim}(t) * a_c(t)/a_c(t+1) * u(t)/u(t+1). \tag{6.6}$$

With these changes we expect to see an increase in the activity simulation performance compared to the original version of the *Activity Dynamics* framework. To compare simulation performances between the "static" network structure and the introduced "dynamic" network structure, we calculate the root-mean-square error (RMSE) of simulated activity and actual empirical activity over the observed 12 months for each of the two approaches. We calculate the RMSE as described in Equation 6.7.

$$RMSE = \sqrt{\frac{\sum_{t=0}^{n} \overline{(\hat{X}(t) - X(t))^2}}{n}}, \tag{6.7}$$

where $\hat{X}(t)$ is the simulated activity of month t, $X(t)$ is the actual empirical activity gained form our data sets of month t and n is the total number of observed months (12 in this work).

7 Results

This chapter contains a description of the results of the activity dynamics simulations for our three conducted experiments, the *Activity Dynamics Scenario* described in Chapter 4, the *Centrality Analysis* explained in Chapter 5 and the *Dynamic Network Structure* depicted in Chapter 6. The interpretation and discussion of these results can be found in Chapter 8.

7.1 Activity Dynamics Scenarios

We studied the impact of six plausible real-world events on online collaboration networks through the implementation of six scenarios explained in Chapter 4. These scenarios were applied to all of our five StackExchange.com networks and all five Semantic MediaWiki networks. Table 7.1 lists the impacts of each scenario on these data sets for the *random* and *informed* approach. All numbers mentioned in Table 7.1 and the following sections represent the relative difference between simulated activity with no manipulations done to the network and simulated activity resulting from the manipulations applied to the network in each of the scenarios respectively at the end of simulations.

Table 7.1: **Absolute Differences to Non-Manipulated Simulated Activity at the End of Simulation (Month 12).** This table lists the differences in activity after 12 months have been simulated for every scenario explained in Section 4, applied to our 10 empirical data sets listed in Section 3.2. The top half shows results for the StackExchange.com networks and the bottom half lists observations for our Semantic MediaWiki data sets. *N* represents the amount of structural changes performed in a particular scenario: 1%, 5% and 10% of existing users and 10%, 30% and 50% of existing collaboration edges are either added or removed form the network, whereas 1, 5 and 10 trolls or incentivized users are added to the network. Results of the *random* and *informed* approach are listed as the relative difference to simulated activity with no manipulations done to the networks.

Scenario	*N*	**Beer-StackExchange**		**Bitcoin-StackExchange**		**Electronics-StackExchange**		**Physics-StackExchange**		**Gaming-StackExchange**	
		informed	*random*	*informed*	*random*	*informed*	*random*	*informed*	*random*	*informed*	*random*
Mass Emigration	1%	−19.1%	−2.1%	−74.1%	−2%	−84.3%	−2.1%	−58.1%	−1.7%	−55%	−1.8%
	5%	−61.7%	−10.6%	−98.9%	−9.6%	−99.8%	−9%	−92.6%	−10.9%	−99.8%	−10.2%
	10%	−95.7%	−19.1%	−100%	−19.4%	−100%	−19.2%	−99.7%	−20.4%	−99.7%	−19.2%
Mass Immigration	1%	+2.1%	+2.1%	+1.2%	+1.1%	+1.8%	+1.1%	+1.1%	+1.1%	+3.4%	+1.1%
	5%	+4.3%	+6.4%	−1.7%	+5.6%	+3.2%	+5.4%	−1.3%	+5.4%	+6.4%	+5.3%
	10%	+4.3%	+12.8%	−8.6%	+11.1%	−2.2%	+10.8%	−7.2%	+10.8%	−6.2%	+10.6%
Breaking Collaborative Ties	10%	−8.5%	−2.1%	−14.3%	−1.1%	−12.5%	−0.5%	−15.1%	−0.6%	−11.4%	−0.3%
	30%	−25.5%	−6.4%	−32.8%	−4.1%	−32.5%	−2.1%	−34.5%	−2.4%	−26.3%	−1.6%
	50%	−48.9%	−10.6%	−54.8%	−9.9%	−52.1%	−4.8%	−53.1%	−5.6%	−37.7%	−3.6%
Establishing New Collaborations	10%	+4.3%	+2.1%	+6.7%	+0.7%	+25.4%	+0.6%	+27.9%	+0.6%	+36.1%	+0.5%
	30%	+17%	+6.4%	+18.6%	+2.5%	+29.1%	+1.7%	+27.8%	+1.8%	+50%	+1.5%
	50%	+17%	+10.6%	+15.8%	+4.1%	+22.3%	+2.9%	+23.2%	+3.1%	+25.8%	+2.5%
Providing Incentives	1	+151.1%	+12.8%	+25.4%	+0.5%	+3.5%	0%	+3.5%	0%	+3.8%	+0.3%
	5	+485.1%	+78.7%	+133.1%	+4.7%	+15.4%	+0.2%	15.9%	+0.2%	+21.7%	+0.2%
	10	+672.3%	+123.4%	+233.3%	+6.4%	+27.2%	+0.6%	+26%	+0.3%	+39.6%	+0.7%
Emergence of Trolls	1	−100%	−100%	−43.8%	−30.9%	−14.3%	−0.2%	−17.7%	−0.2%	−10.7%	−0.7%
	5	−100%	−100%	−95.2%	−100%	−54%	−0.8%	−62.2%	−100%	−43.4%	−3.4%
	10	−100%	−100%	−100%	−100%	−78.9%	−1.6%	−85.8%	−1.2%	−68.1%	−6.9%

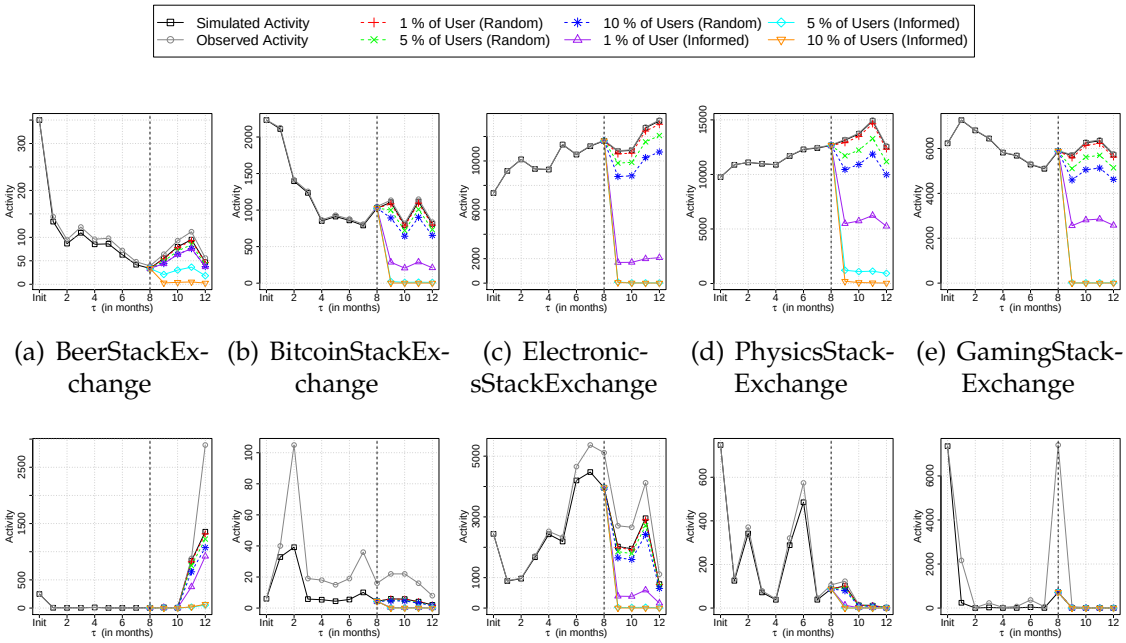
Scenario	*N*	**ComplexOperations**		**BioInformatics**		**NeuroLex**		**DotaWiki**		**PracticalPlants**	
		informed	*random*	*informed*	*random*	*informed*	*random*	*informed*	*random*	*informed*	*random*
Mass Emigration	1%	−32.2%	−2.8%	−100%	0%	−79.7%	−1.8%	−50%	0%	−100%	0%
	5%	−96.1%	−9.6%	−100%	0%	−99.1%	−8.1%	−100%	0%	−100%	0%
	10%	−95.1%	−20.8%	−100%	0%	−100%	−18%	−100%	0%	−100%	0%
Mass Immigration	1%	+0.9%	+1.4%	0%	0%	+1.5%	+1.1%	0%	0%	0%	0%
	5%	+2.1%	+6.6%	0%	0%	+5.8%	+5.6%	0%	0%	0%	0%
	10%	+1.8%	+13.5%	0%	0%	+8.5%	+11.2%	0%	0%	0%	0%
Breaking Collaborative Ties	10%	−21.2%	−3%	0%	0%	−9.6%	−0.9%	0%	0%	0%	0%
	30%	−25.8%	−16.5%	−50%	0%	−27%	−5.3%	−50%	0%	0%	0%
	50%	−96.5%	−31.1%	−100%	−50%	−42.7%	−12.9%	−50%	0%	−100%	0%
Establishing New Collaborations	10%	+21.4%	+2.7%	0%	0%	+1.4%	+0.8%	0%	0%	0%	0%
	30%	+24.7%	+7.8%	0%	0%	+0.6%	+2.4%	0%	0%	0%	0%
	50%	+22.4%	+13.8%	0%	0%	+3.9%	+4.1%	0%	0%	0%	0%
Providing Incentives	1	+16.2%	+7.1%	+1400%	+200%	+9%	+0.4%	+3950%	+50%	+11200%	0%
	5	+18.2%	+14.4%	+2650%	+700%	+42.8%	+1.9%	+10700%	+400%	+20600%	+200%
	10	+20.1%	+15.6%	+3000%	+1050%	+66.7%	+3.7%	+15400%	+1350%	+27500%	+1100%
Emergence of Trolls	1	−100%	−100%	−100%	−100%	−84.3%	−0.5%	−100%	−100%	−100%	−100%
	5	−100%	−100%	−100%	−100%	−100%	−2.5%	−100%	−100%	−100%	−100%
	10	−100%	−100%	−100%	−100%	−100%	−100%	−100%	−100%	−100%	−100%

7.1.1 Mass Emigration

We present the resulting levels of activity for this scenario, for each of the StackExchange.com networks and each of the Semantic MediaWiki networks, in Figure 7.1. The removal of existing users from the collaboration network negatively influences activity on all empirical data sets, regardless of the applied approach (*random* or *informed*) and the amount of existing users removed from the networks (1%, 5% and 10% of existing users).

As Table 7.1 lists, removing 1% of existing users results in a loss of 1.7% to 2.1% for the StackExchange.com networks and 0% to 2.7% for the Semantic MediaWiki networks. When randomly removing 5% of all existing users of the networks, these numbers range from 9% to 10.9% and 0% to 9.6%. The removal of 10% of existing users from the collaboration network resulted in a decrease of activity ranging from 19.1% to 20.4% for the StackExchange.com data sets and from 0% to 20.8% for the Semantic MediaWiki data sets relative to the non-manipulated simulated activity resulting from the empirical networks.

The *informed* approach has a stronger impact on overall activity and therefore harms the networks more effectively than the *random* approach. Specifically removing 1% of existing users already reduces overall activity in a range of 19.1% to 84.3% for our StackExchange.com networks and 32.2% to 100% for our Semantic MediaWiki networks. By specifically removing 5% of all existing nodes, the decrease of activity ranges from 61.7% to 99.8% for the StackExchange.com data sets and from 96.1% to 100% for our Semantic MediaWiki data sets. With the removal of 10% of users from the collaboration network, all Semantic MediaWiki networks show zero activity at the end of the simulation (100% decrease), leaving them in a "dead" state. The relative decrease ranges between 95.7% and 100% for our StackExchange.com networks.



(a) BeerStackExchange (b) BitcoinStackExchange (c) ElectronicsStackExchange (d) PhysicsStackExchange (e) GamingStackExchange

(f) ComplexOperations (g) BioInformatics (h) NeuroLex (i) DotaWiki (j) PracticalPlants

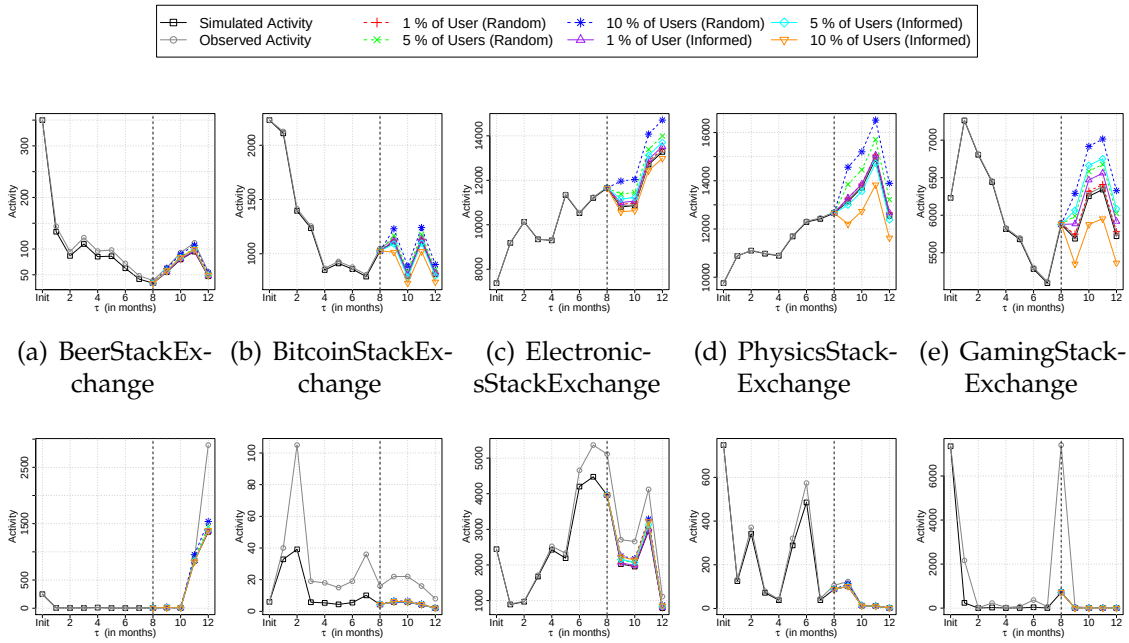
Figure 7.1: **Results of the *Mass Emigration* Scenario.** This figure illustrates the simulation results for the *Mass Emigration* scenario applied to our StackExchange.com data sets (top) and our Semantic MediaWiki data sets (bottom). The gray lines with circles represent empirically observed activities of the corresponding data sets, while the black lines with squares represent the non-manipulated simulated activities of the original *Activity Dynamics* framework. The remaining lines represent the simulation results of our different approaches of the scenario. Each approach is simulated for 4 months, starting at month 8 (indicated by the vertical dashed line). We observe a decrease in activity for all used empirical data sets if users leave the networks, which is especially affecting the network's overall activity when we specifically select the high degree nodes (*informed* approach) and remove them from the network.

7.1.2 Mass Immigration

This scenarios is the exact counterpart to the *Mass Emigration* scenario. Whenever users are randomly added to the collaboration network, a proportional increase in activity can be seen (see Figure 7.2). Similar observations have been made for the *informed* approach and the Semantic MediaWiki networks. However, when we added more than the additional 5% to the StackExchange.com networks and linked them with high degree users (*informed*), activity decreases. The numbers for both approaches and the different amounts of users added to our empirical networks (1%, 5% and 10% of existing users) are listed in Table 7.1.

The following observations have been made for this scenario: When adding the total amount of 1% of existing users and randomly connecting them with existing ones, the gain in activity ranges from 1.1% to 2.1% for all StackExchange.com networks and from 0% to 1.1% for all Semantic MediaWiki networks. When we add the amount of 5% of existing users, the increase of activity ranges between 5.3% and 6.4% for StackExchange.com data sets and between 0% and 6.6% for the Semantic MediaWiki data sets. By adding the amount of 10% of users, activity is lifted in the range of 10.6% to 12.8% for the StackExchange.com networks and between 0% and 13.5% for the Semantic MediaWiki networks.

Again, the *informed* approach had a more decisive impact on simulation results compared to the *random* approach. By adding the amount of 1% of existing users, activity sees a gain between 1.1% and 3.4% for our StackExchange.com networks and between 0% and 1.5% for our Semantic MediaWiki networks. For the approach, where the amount of 5% of existing users were added to the collaboration network, these numbers are -1.3% to 6.4% and 0% to 5.8%. Note the decrease of activity for two of the StackExchange.com networks (illustrated in Figure 7.2(b) and Figure 7.2(d)). When we add the amount of 10% of existing users to our empirical networks, the StackExchange.com data sets lost or gained activity in a range from -8.6% and 4.3% and the Semantic MediaWiki networks gained activity in the range from 0% to 8.5%.



(a) BeerStackExchange (b) BitcoinStackExchange (c) ElectronicsStackExchange (d) PhysicsStackExchange (e) GamingStackExchange

(f) ComplexOperations (g) BioInformatics (h) NeuroLex (i) DotaWiki (j) PracticalPlants

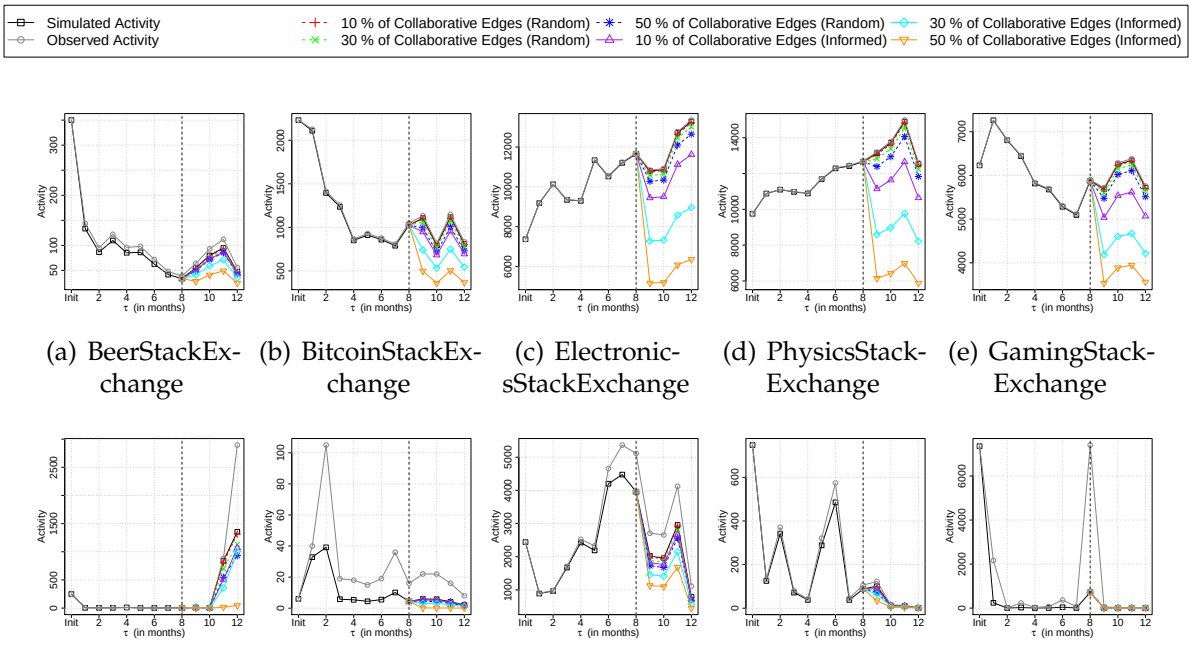
Figure 7.2: **Results of the *Mass Immigration* scenario.** This figure illustrates the simulation results for the *Mass Immigration* scenario applied to our StackExchange.com data sets (top) and our Semantic MediaWiki data sets (bottom). The gray lines with circles represent empirically observed activities of the corresponding data sets, while the black lines with squares represent the non-manipulated simulated activities of the original *Activity Dynamics* framework. The remaining lines represent the simulation results of our different approaches of the scenario. Each approach is simulated for 4 months, starting at month 8 (indicated by the vertical dashed line). We observe an increase of activity whenever we applied our *random* approach. However, the *informed* approach reveals a possible decrease of activity for all the StackExchange.com networks (top) if the amount of newly added users is high enough. In contrast to that, activity increases in all Semantic MediaWiki networks for the *informed* approach.

7.1.3 Breaking Collaborative Ties

The results of this scenario are similar to the ones observed in the *Mass Emigration* scenario where activity decreases in each approach, as illustrated by Figure 7.3. In general, we observed a decrease of overall activity that is proportional to the amount of collaboration edges removed from the network. Exact numbers of the changes in overall activity levels for each approach (*random* and *informed*) and each amount of collaboration edges removed from the network (10%, 30% and 50% of existing users) can be found in Table 7.1.

When we remove 10% of existing collaboration edges, the StackExchange.com networks show a reduction of activity ranging from 0.3% to 2.1% and the Semantic MediaWiki networks a reduction ranging from 0% to 3%. The removal of 30% of existing edges from the network resulted in 1.6% to 6.4% decrease of activity for the StackExchange.com data sets and 0% - 16.5% for the Semantic MediaWiki data sets. Randomly removing 50% of existing collaboration edges leads to a decrease of activity in the range from 3.6% to 10.6% for StackExchange.com networks and from 0% to 31.1% for the Semantic MediaWiki networks.

Similar to the other scenarios, the *informed* approach harms the network's overall activity more effectively than the *random* approach of this scenario. The loss of activity by the specific removal of 1% of edges between high degree users varies between 8.5% and 15.1% for the StackExchange.com data sets and between 0% and 21.2% for the Semantic MediaWiki data sets. By specifically removing 30% of existing collaboration edges, overall levels of activity of the StackExchange.com data sets were decrease by 26.3% to 34.5% and overall activity of the Semantic MediaWiki networks by 0% to 50%. Again, when we remove 50% of existing collaboration edges, the network is harmed the most: The decrease of activity ranges from 37.7% to 54.8% for our StackExchange.com networks and from 42.7% to 100% for our Semantic MediaWiki networks.



(a) BeerStackExchange (b) BitcoinStackExchange (c) ElectronicsStackExchange (d) PhysicsStackExchange (e) GamingStackExchange

(f) ComplexOperations (g) BioInformatics (h) NeuroLex (i) DotaWiki (j) PracticalPlants

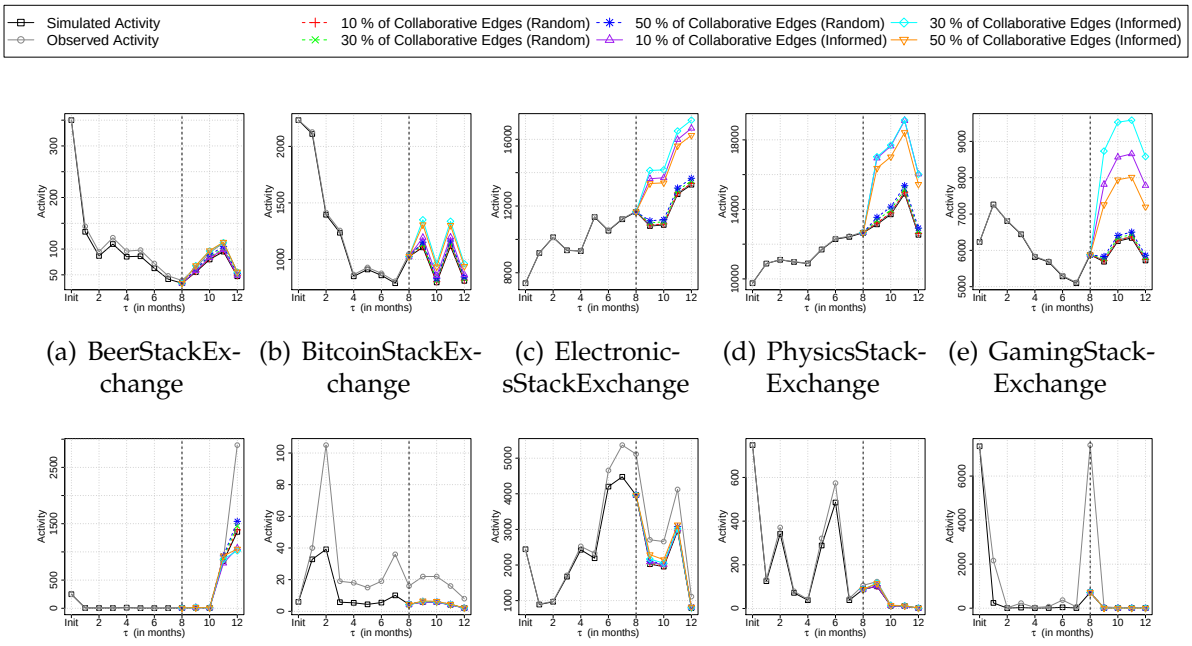
Figure 7.3: **Results of the *Breaking Collaborative Ties* Scenario.** This figure illustrates the simulation results for the *Breaking Collaborative Ties* scenario applied to our StackExchange.com data sets (top) and our Semantic MediaWiki data sets (bottom). The gray lines with circles represent empirically observed activities of the corresponding data sets, while the black lines with squares represent the non-manipulated simulated activities of the original *Activity Dynamics* framework. The remaining lines represent the simulation results of our different approaches of the scenario. Each approach is simulated for 4 months, starting at month 8 (indicated by the vertical dashed line). When we remove any amount of collaboration edges from the collaboration networks, it simultaneously equals a reduction of overall activity. However, even when we randomly remove 50% of all edges, activity does not reach zero and keeps almost the same level of activity compared to non-manipulated simulations.

7.1.4 Establishing New Collaborations

We added new connections between existing users within the collaboration networks for this scenario. Our results show that the levels of activity are increased for all ten empirical data sets. However, the level of overall activity is not affected as much as one would expect when adding half the amount of already existing edges between *randomly* selected users. Our *informed* approach shows that activity within networks is increased more significantly. Figure 7.4 depicts the resulting plots for all StackExchange.com data sets and all Semantic MediaWiki data sets.

By *randomly* adding 10% of the amount of existing edges, we pushed activity by an increase of 0.5% to 2.1% for StackExchange.com networks, when we add the amount of 30% of existing edges activity increases by 1.5% to 6.4% and when adding half of the amount of existing edges (50%), a gain of 2.5% to 10.6% of activity is observed. For the Semantic MediaWiki data sets the increase in activity is just as small. When we add 10% of existing edges, activity is pushed by 0% to 2.7%, when adding 30% of existing edges activity is increased by 0% to 7.8% and when we add half of the amount of already existing collaboration edges, activity for Semantic MediaWiki networks is pushed by a range of 0% to 13.8%.

Our *informed* approach increased the overall activity within the network more than the *random* one. When we, in particular, take a look on the StackExchange.com data sets, activity increases between 4.3% to 25.4% for 10% added edges, between 17% to 50% for 30% added edges and between 15.8% to 25.8% for half of the amount of existing edges. When we take a look on the Semantic MediaWiki data sets, these values are between 0% and 21.4% for 10% added edges, between 0% and 24.7% for 30% added edges and between 0% and 22.4% for 50% added edges.



(a) BeerStackExchange (b) BitcoinStackExchange (c) ElectronicsStackExchange (d) PhysicsStackExchange (e) GamingStackExchange

(f) ComplexOperations (g) BioInformatics (h) NeuroLex (i) DotaWiki (j) PracticalPlants

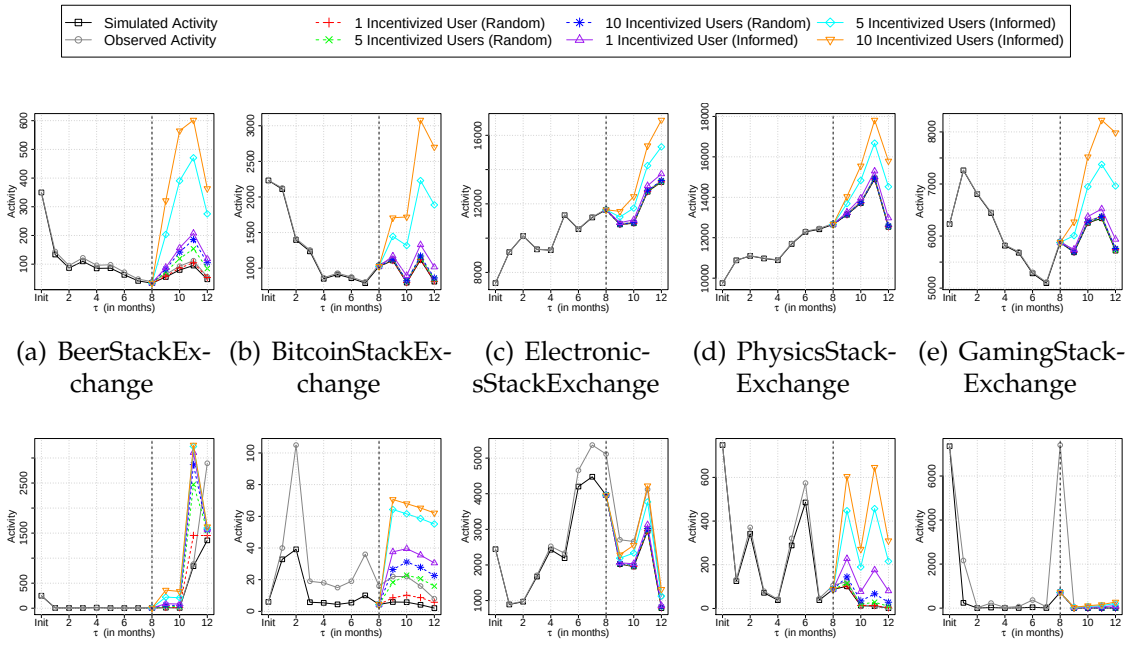
Figure 7.4: **Results of the *Establishing New Collaborations* Scenario.** This figure illustrates the simulation results for the *Establishing New Collaborations* scenario applied to our StackExchange.com data sets (top) and our Semantic MediaWiki data sets (bottom). The gray lines with circles represent empirically observed activities of the corresponding data sets, while the black lines with squares represent the non-manipulated simulated activities of the original *Activity Dynamics* framework. The remaining lines represent the simulation results of our different approaches of the scenario. Each approach is simulated for 4 months, starting at month 8 (indicated by the vertical dashed line). A noticeable increase in the overall levels of activity for all the StackExchange.com data sets (top) can be observed for the *informed* approach, whereas our *random* approach is not notably affecting the network's overall activity. Adding edges is nearly indistinguishable from the non-manipulated simulated activity for all Semantic MediaWiki data sets (bottom).

7.1.5 Providing Incentives

According to our results for the five StackExchange.com data sets and the five Semantic MediaWiki networks depicted in Figure 7.5, providing incentives increases the overall activity within collaboration networks, with exceptions in the following networks: The *ElectronicsStackExchange* (Figure 7.5(c)), *PhysicsStackExchange* (Figure 7.5(d)) as well as the *PracticalPlants* (Figure 7.5(j)) networks did not experience an increase in overall activity when 1 randomly incentivized user was added to the networks. All specific numbers for the *random* and *informed* approaches and all amounts of added incetivized users (1, 5 and 10) are listed in Table 7.1.

In the *random* approach, were we introduced 1 incentivized user to the networks, we increased the overall activity between 0% and 12.8% for the StackExchange.com networks and between 0% and 200% for the Semantic MediaWiki networks. We reached a higher gain in activity by randomly selecting 5 existing users, where the StackExchange.com networks show an increase in activity between 0.2% and 78.7% and all our Semantic MediaWiki networks show an increase in activity between 1.9% to 700%. Even better results are reached when *randomly* adding 10 incentivized users, where the activity of the StackExchange.com networks could be increased by 0.3% to 123.4% and the activity of the Semantic MediaWiki networks between 3.7% and 1050%.

Our *informed* approach increased activity even more. Between 3.5% and 151.1% for our StackExchange.com networks and 9% and 11200% for the Semantic MediaWiki networks could be reached by only adding 1 incentivized user with the highest degree among all existing users in the collaboration network. When specifically adding 5 incentivzed users, we reached a gain between 15.4% and 485.1% for StackExchange.com data sets and between 18.2% and 20600% for the Semantic MediaWiki data sets. By adding a total amount of 10 incentivized users, the activity could be pushed between 26% and 672.3% for the StackExchange.com networks and between 20.1% and 27500% for the Semantic MediaWiki networks.



(a) BeerStackExchange (b) BitcoinStackExchange (c) ElectronicsStackExchange (d) PhysicsStackExchange (e) GamingStackExchange

(f) ComplexOperations (g) BioInformatics (h) NeuroLex (i) DotaWiki (j) PracticalPlants

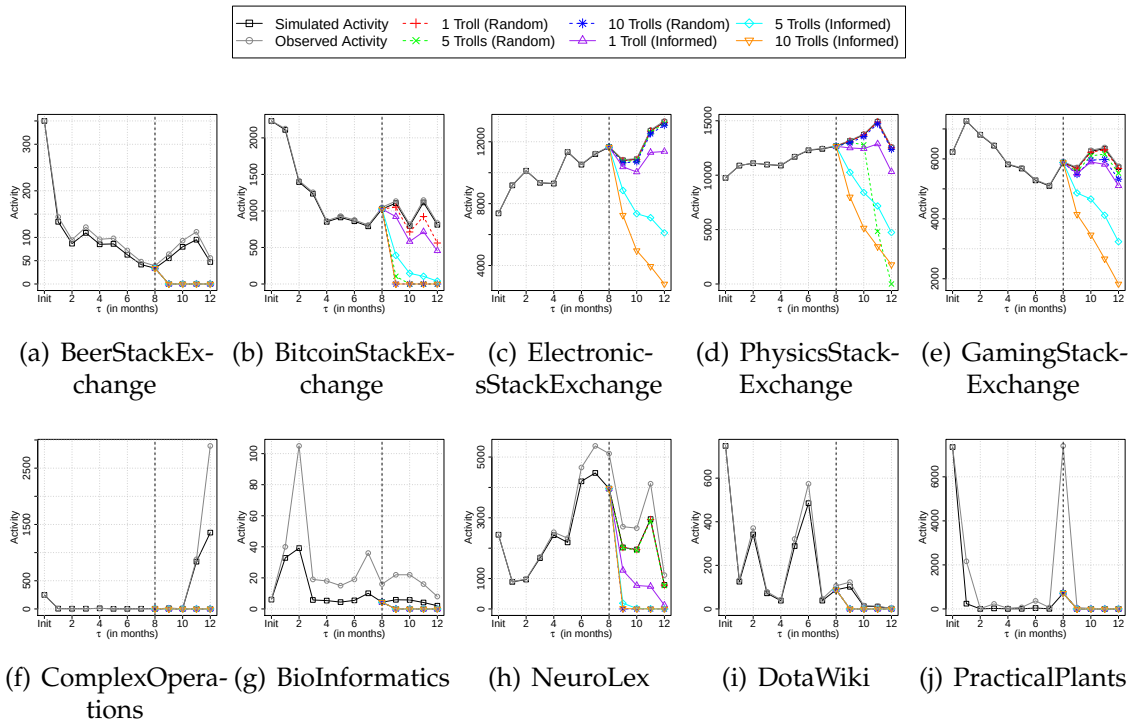
Figure 7.5: **Results of the *Providing Incentives* Scenario.** This figure illustrates the simulation results for the *Providing Incentives* scenario applied to our StackExchange.com data sets (top) and our Semantic MediaWiki data sets (bottom). The gray lines with circles represent empirically observed activities of the corresponding data sets, while the black lines with squares represent the non-manipulated simulated activities of the original *Activity Dynamics* framework. The remaining lines represent the simulation results of our different approaches of the scenario. Each approach is simulated for 4 months, starting at month 8 (indicated by the vertical dashed line). Providing incentives increases the overall levels of activity for all our data sets, except for *ElectronicsStackExchange* (c), *PhysicsStackExchange* (d) and *PracticalPlants* (j), where the introduction of one *randomly* added incentivized user had only a minimal influence on overall activity.

7.1.6 Emergence of Trolls

We present the results for the introduction of trolls to our ten empirical collaboration networks in Figure 7.6. In all our smaller networks, the introduction of only one single troll—connected either *randomly* or *informed* to existing users—reduces the activity to zero, leaving the network in a "dead" state for the rest of the simulation (see Figures 7.6(a), 7.6(f), 7.6(g), 7.6(i) and 7.6(j)). In more active collaboration networks, the activity is decreased proportional to the number of trolls added during the *random* and *informed* approaches. The Exact results for our *random* and *informed* approaches and for all amounts of introduced trolls (1, 5, 10) can be found in Table 7.1.

When introducing trolls to StackExchange.com networks and connecting them through the *random* approach, results depend on the total amount of trolls added. If we only add 1 troll to the networks, activity declines by 0.2% to 100%. When adding 5 trolls, activity is decreased between 0.8% and 100%. Finally, when adding 10 trolls to our StackExchange.com networks, activity is reduced between 1.2% and 100%. We observe a similar outcome for all Semantic MediaWiki networks. Adding 1 troll to these networks decreases activity between 0.5% to 100%. Adding a total of 5 trolls decreases activity between 2.5% and 100%. When we add a total of 10 trolls to these networks, activity declines by 100% to zero overall activity, ultimately resulting in "dead" networks.

Again, our *informed* approach significantly damages all of our ten empirical networks. All StackExchange.com data sets see a decrease of activity that ranges from 10.7% to 100% when we add 1 troll to the networks. By adding 5 trolls, we reached a decrease of activity between 43.4% and 100% compared to non-manipulated activity. The total amount of 10 trolls added to our StackExchange.com networks leads to an activity decrease between 68.1% and 100%. Our Semantic MediaWiki networks, which are smaller in size, had more difficulties in dealing with trolls. When we added 1 troll to these networks, activity declined between 84.3% and 100%. The amount of 5 and 10 introduced trolls stopped overall activity of our Semantic MediaWiki networks for the rest of simulations as the activity decreased by 100%.



(a) BeerStackExchange (b) BitcoinStackExchange (c) ElectronicsStackExchange (d) PhysicsStackExchange (e) GamingStackExchange

(f) ComplexOperations (g) BioInformatics (h) NeuroLex (i) DotaWiki (j) PracticalPlants

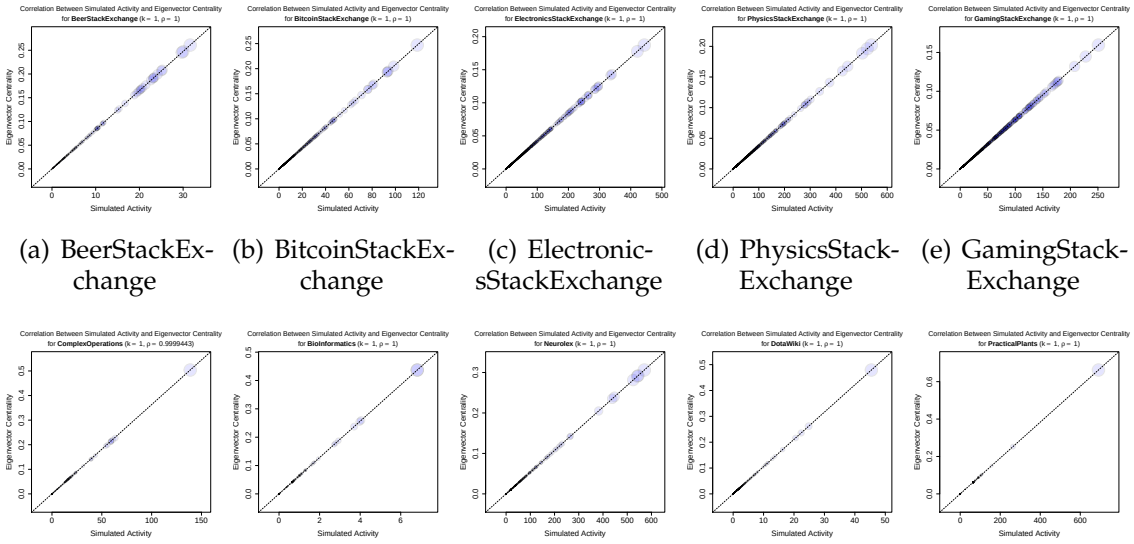
Figure 7.6: **Results of the *Emergence of Trolls* scenario.** This figure illustrates the simulation results for the *Emergence of Trolls* scenario applied to our StackExchange.com data sets (top) and our Semantic MediaWiki data sets (bottom). The gray lines with circles represent empirically observed activities of the corresponding data sets, while the black lines with squares represent the non-manipulated simulated activities of the original *Activity Dynamics* framework. The remaining lines represent the simulation results of our different approaches of the scenario. Each approach is simulated for 4 months, starting at month 8 (indicated by the vertical dashed line). The more trolls we introduce, the harder it becomes for all empirical data sets to maintain (wanted) high levels of activity.

7.2 Centrality Analysis

This section covers all the results obtained from our *Centrality Analysis* experiment explained in Chapter 5. As we observed similar results for all our empirical data set, we have only visualized results where the corresponding correlation is the maximum out of all possible values of k for each of our StackExchange.com data sets and each of our Semantic MediaWiki data sets. However, we list the calculated correlation coefficients for all empirical data sets in Table 7.2 at the end of this section. The following paragraphs describe the results for each approach conducted as part of this experiment.

Eigenvector Centrality. The Pearson Correlation Coefficient between simulated activity and eigenvector centrality is exactly 1 for all our empirical data sets and all values of k (1, 2 and 3).

Hence, we observed a linear correlation between these two values and potentially uncovered a *limitation* of the *Activity Dynamics* framework. We will discuss this incident later on in Chapter 8 *Discussion*. Figure 7.7 illustrates the linear correlation for all StackExchange.com networks and all Semantic MediaWiki networks for $k = 1$.



(a) BeerStackExchange (b) BitcoinStackExchange (c) ElectronicsStackExchange (d) PhysicsStackExchange (e) GamingStackExchange

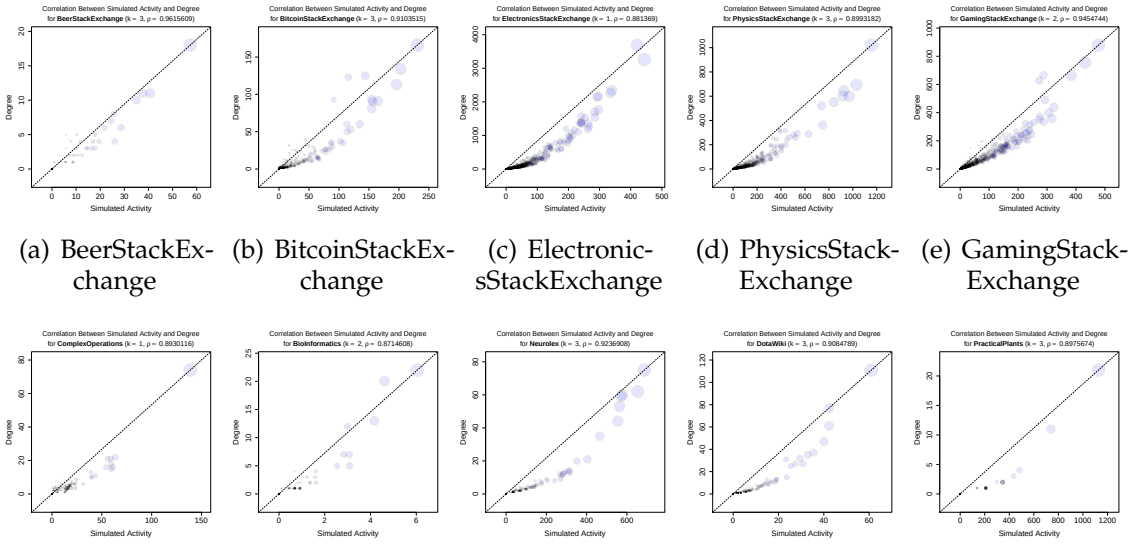
(f) ComplexOperations (g) BioInformatics (h) NeuroLex (i) DotaWiki (j) PracticalPlants

Figure 7.7: **Correlation Between Eigenvector Centrality and Simulated Activity for All Empirical Networks.** This figure illustrates the correlation between simulated activity and eigenvector centrality for all StackExchange.com data sets (top) and all Semantic MediaWiki data sets (bottom). Blue circles represent users and circle sizes show eigenvector centrality. The value of k states the required amount of interaction between two users in order to connect them with a collaboration edge. ρ is the calculated Pearson Correlation Coefficient. Simulated activity and eigenvector centrality of nodes correlate linearly as $\rho = 1$ for all empirical networks, potentially uncovering a *limitation* of the *Activity Dynamics* framework.

Degree Centrality. The degree of nodes correlates strongly with the simulated activity and gets either closer to 1 (total correlation) or closer to 0 whenever the required amount of interaction between two users in order to connect them with a collaboration edge is set higher ($k = 2$ and $k = 3$). When we construct the network in the original way of the *Activity Dynamics* framework ($k = 1$), the Pearson Correlation Coefficients are 0.942 for *BeerStackExchange*, 0.898 for *BitcoinStackExchange*, 0.881 for *ElectronicsStackExchange*, 0.886 for *PhysicsStackExchange* and 0.940 for the *GamingStackExchange* network. Similar numbers have been observed for the Semantic MediaWiki networks. The degree centrality of users within *ComplexOperations* correlates with simulated activity with 0.893, for *BioInformatics* with 0.859, for *NeuroLex* with 0.875, for *DotaWiki* with 0.889 and for *PracticalPlants* with 0.811.

The correlation coefficients are higher for a higher number of k. So when we set the needed amount of interactions to create a collaboration edge between to users to $k = 2$ and further to $k = 3$, we get 0.952 and 0.962 for the *BeerStackExchange* network, 0.907 and 0.910 for *BitcoinStackExchange*, 0.878 and 0.880 for *ElectronicsStackExchange*, 0.892 and 0.899 for *PhysicsStackExchange* and 0.945 and 0.943 for *GamingStackExchange*. Note that if $k = 3$ the *GamingStackExchange* network showed different behavior to the other StackExchange.com networks by obtaining a smaller correlation coefficient compared to $k = 2$. Similar to this, we have our Semantic MediaWiki networks with 0.812 $(k = 2)$ and 0.806 $(k = 3)$ for *ComplexOperations*, 0.871 and 0.870 for *BioInformatics*, 0.882 and 0.924 for *NeuroLex*, 0.862 and 0.908 for *DotaWiki* and 0.856 an 0.989 for *PracticalPlants*.

Figure 7.8 illustrates the maximum correlation between simulated activity and degree among all values of k (1, 2 or 3) for each of our empirical data sets.



(a) BeerStackExchange (b) BitcoinStackExchange (c) ElectronicsStackExchange (d) PhysicsStackExchange (e) GamingStackExchange

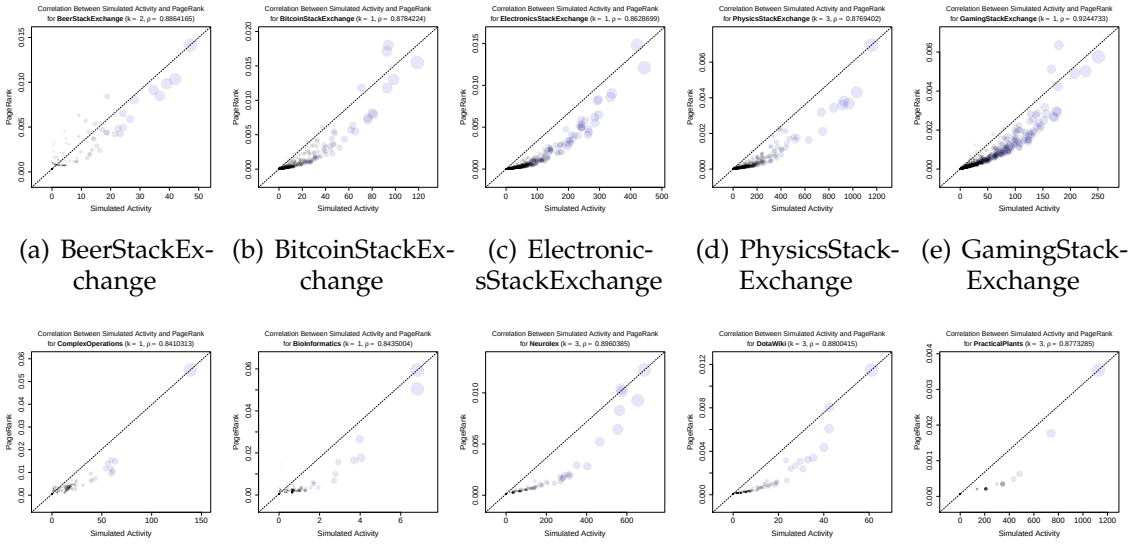
(f) ComplexOperations (g) BioInformatics (h) NeuroLex (i) DotaWiki (j) PracticalPlants

Figure 7.8: **Highest Correlation Between Degree and Simulated Activity for All Empirical Networks.** This figure illustrates the correlation between simulated activity and degree for all StackExchange.com data sets (top) and all Semantic MediaWiki data sets (bottom). Blue circles represent users and circle sizes show degree centrality. The value of k states the required amount of interaction between two users in order to connect them with a collaboration edge. ρ is the calculated Pearson Correlation Coefficient. Simulated activity and degree of nodes correlate strongly. Note that the value of k is different for our networks, as we only illustrate the plots with the highest correlation among all values of k (1, 2 or 3).

PageRank Centrality. The correlation between Google's PageRank and simulated activity resulting from the *Activity Dynamics* framework is different for each network and each k. The calculated Pearson Correlation Coefficient for $k = 1$ and *BeerStackExchange* is 0.882. This raises to 0.886 if we set $k = 2$ and decreases to 0.879 if $k = 3$. In the case of the *BitcoinStackExchange* network, the coefficients decrease from 0.878 with $k = 1$ to 0.875 with $k = 2$ and to 0.867 with $k = 3$, so they decrease with a higher k. The correlations between PageRank and the simulated activity of the *ElectronicStackExchange* network is similar to this. Here we calculated a coefficient of 0.863 $(k = 1)$, 0.855 $(k = 2)$ and 0.854 $(k = 3)$. With the *PhysicsStackExchange* network, the values increase with a higher k, so with $k = 1$ we obtained 0.867, with $k = 2$ we calculated 0.870 and with $k = 3$ we got 0.877. This is the only occurrence of a constant increase in correlation between simulated Activity and PageRank among our StackExchange.com data sets. The simulated activity within the *GamingStackExchange* network also correlates strongly with PageRank centrality, but again, coefficients are decreasing with a higher k so we get 0.924 $(k = 1)$, 0.922 $(k = 2$ and 0.913 $(k = 3)$.

The resulting correlation coefficients of the Semantic MediaWiki data sets are equally unpredictable in behavior when we change k. The simulated activity in the *ComplexOperations* network correlates strongly with PageRank at 0.841 and $k = 1$. However, this correlation decreases to 0.699 if we set $k = 2$ and to 0.649 if $k = 3$. Similar observations can be made with the *BioInformatics* network where we calculated 0.843 for $k = 1$, 0.839 for $k = 2$ and 0.781 for $k = 3$. The trend of correlations is different for the *NeuroLex* network, where the correlation first decreases and then increases again, depending on k. We obtained 0.839 $(k = 1)$, 0.825 $(k = 2)$ and 0.896 $(k = 3)$ for this network. The correlations of *DotaWiki* evolve similar to *NeuroLex* where with $k = 1$ we get 0.879. This decreases to 0.835 for $k = 2$ and increases again to 0.880 if we set $k = 3$. A constant increase of correlations subject to k can be observed with the *PracticalPlants* network. Here we received a correlation coefficient of 0.796 for $k = 1$, 0.836 for $k = 2$ and 0.877 for $k = 3$.

We show the maximum correlation of PageRank and simulated activity among all possible values of k for each data set in Figure 7.9.



(a) BeerStackExchange (b) BitcoinStackExchange (c) ElectronicsStackExchange (d) PhysicsStackExchange (e) GamingStackExchange

(f) ComplexOperations (g) BioInformatics (h) NeuroLex (i) DotaWiki (j) PracticalPlants

Figure 7.9: **Highest Correlation Between PageRank and Simulated Activity for All Empirical Networks.** This figure illustrates the correlation between simulated activity and PageRank for all StackExchange.com data sets (top) and all Semantic MediaWiki data sets (bottom). Blue circles represent users and circle sizes show degree centrality. The value of k states the required amount of interaction between two users in order to connect them with a collaboration edge. ρ is the calculated Pearson Correlation Coefficient. Simulated activity and PageRank of nodes correlate strongly. Note that the value of k is different for our networks, as we only illustrate the plots with the highest correlation among all values of k (1, 2 or 3).

Table 7.2: **Correlation Coefficients Between Simulated Activity and Centrality Measures.** This table lists the calculated correlation coefficients for the *Centrality Analysis* experiment. Correlation coefficients are listed for $k = 1$ to $k = 3$ and all used empirical data sets.

Data Set	**Centrality Measure**	$k = 1$	$k = 2$	$k = 3$
BeerStackExchange	Degree	0.942	0.952	0.962
	Eigenvector Centrality	1	1	1
	PageRank	0.882	0.886	0.879
BitcoinStackExchange	Degree	0.898	0.907	0.910
	Eigenvector Centrality	1	1	1
	PageRank	0.878	0.875	0.867
ElectronicsStackExchange	Degree	0.881	0.878	0.880
	Eigenvector Centrality	1	1	1
	PageRank	0.863	0.855	0.854
PhysicsStackExchange	Degree	0.886	0.892	0.899
	Eigenvector Centrality	1	1	1
	PageRank	0.867	0.870	0.877
GamingStackExchange	Degree	0.940	0.945	0.943
	Eigenvector Centrality	1	1	1
	PageRank	0.924	0.922	0.913
ComplexOperations	Degree	0.893	0.812	0.806
	Eigenvector Centrality	1	1	1
	PageRank	0.841	0.699	0.649
BioInformatics	Degree	0.859	0.871	0.870
	Eigenvector Centrality	1	1	1
	PageRank	0.843	0.839	0.781
NeuroLex	Degree	0.875	0.882	0.924
	Eigenvector Centrality	1	1	1
	PageRank	0.839	0.825	0.896
DotaWiki	Degree	0.889	0.862	0.908
	Eigenvector Centrality	1	1	1
	PageRank	0.879	0.835	0.880
PracticalPlants	Degree	0.811	0.856	0.898
	Eigenvector Centrality	1	1	1
	PageRank	0.796	0.836	0.877

7.3 Dynamic Network Structure

This section covers the results of our *Dynamic Network Structure* experiment explained in Chapter 6. Using an underlying dynamic network structure for simulation of activity with the *Activity Dynamics* framework, significantly increases the performance of simulations for all empirical data sets, except for the *PracticalPlants* network among our Semantic MediaWiki data sets.

We list the differences in user numbers, number of collaboration edges and the variances of parameters needed for simulation for each of the observed months for one StackExchange.com data set and one Semantic MediaWiki data set in Table 7.4. Due to the fact that all networks showed similar results for this experiments, we only describe changes in the number of users, collaboration edges, and the difference between simulated and empirical activity in the following paragraphs and do not list them in particular tables.

As Figure 7.10 depicts for all our empirical networks, the performance of activity simulation is very accurate, as all up and down trends of empirical activity are correctly simulated for the StackExchange.com networks (see Figure 7.10(a) to Figure 7.10(e)). Note, that in these networks, a large amount of activity comes from newly joined users at the beginning of each month (visualized by the dashed green lines). This can especially be observed in Figure 7.10(c), Figure 7.10(d) and Figure 7.10(e).

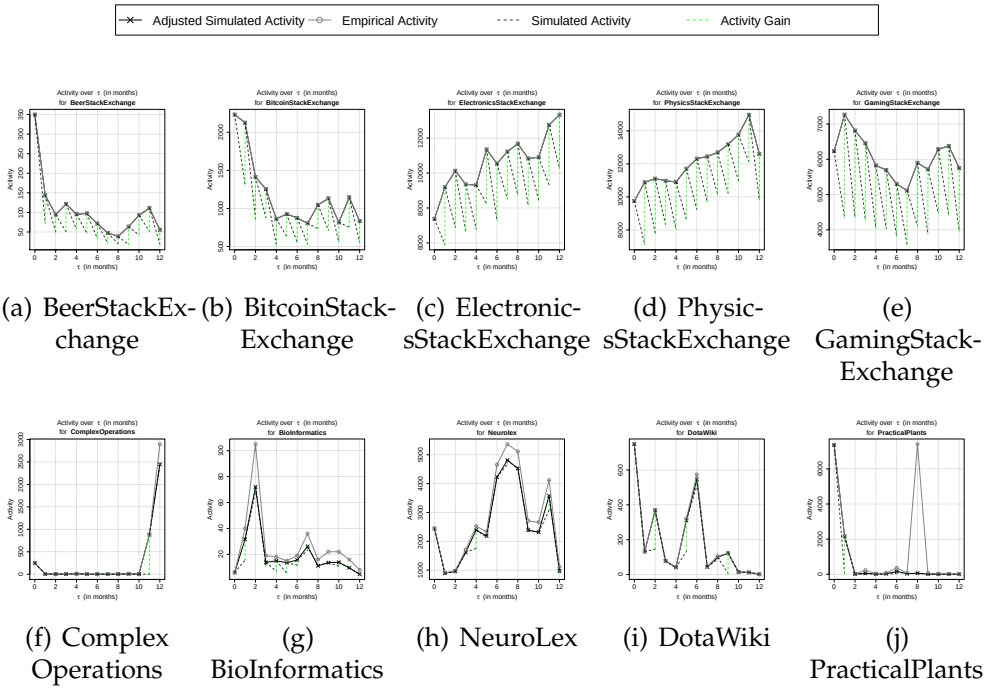
In contrast to the StackExchange.com networks, the *Activity Dynamics* model did not simulate activity as accurately for all Semantic MediaWiki networks (see Figure 7.10(f) to Figure 7.10(j)). Furthermore, the amount of activity coming from newly joined users at the beginning of each month is not as significant compared to the StackExchange.com data sets. Sudden high gains in activity were not simulated correctly as seen in Figure 7.10(g) and especially in Figure 7.10(j).

The root-mean-square error (RMSE) of the simulated activity and actual empirical activity of the *BeerStackExchange* network was decreased by 9.64 compared to the original "static" approach. For the *BitcoinStackExchange* network, this decrease was even higher and the RMSE could be reduced by 18.09. The RMSE of the *ElectronicsStackExchange* network was reduced by

26.46, of the *PhysicsStackExchange* by 32.45 and of the *GamingStackExchange* by 19.95. Similar observations were made for the Semantic MediaWiki networks. Our "dynamic" approach reduced the RMSE of *ComplexOperations* by 302.75 and of *BioInformatics* by 11.60. Further, we see a decrease of the RMSE by 275.16 for the *NeuroLex* network and by 20.32 for the *DotaWiki* network, whereas the PracticalPlants network is the aforementioned exception. Here, we increased the RMSE by 104.69 with our "dynamic" approach. All exact values can be found in Table 7.3.

Table 7.3: **RMSE of the Resulting Simulated Activity and Actual Empirical Activity.** This table lists the RMSE of the original version of the *Activity Dynamics* framework where the underlying network structure is considered to be static, and the RMSE of the improved approach, where the underlying network structure and parameter calculation are dynamic and change over time. Column *Absolute Change* shows the absolute differences between these two RMSE. Performances of all simulations for our empirical networks have improved as the RMSE shrinks, except for PracticalPlants where the RMSE was increased.

Datasets	**RMSE (static)**	**RMSE (dynamic)**	**Absolute Change**
BeerStackExchange	10.22	0.58	−9.64
BitcoinStackExchange	20.90	2.81	−18.09
ElectronicsStackExchange	34.48	8.02	−26.46
PhysicsStackExchange	44.80	12.35	−32.45
GamingStackExchange	23.89	3.94	−19.95
ComplexOperations	427.27	124.52	−302.75
BioInformatics	22.40	10.81	−11.60
NeuroLex	606.76	331.60	−275.16
DotaWiki	28.82	8.50	−20.32
PracticalPlants	1933.99	2038.68	+104.69



(a) BeerStackExchange (b) BitcoinStackExchange (c) ElectronicsStackExchange (d) PhysicsStackExchange (e) GamingStackExchange

(f) Complex Operations (g) BioInformatics (h) NeuroLex (i) DotaWiki (j) PracticalPlants

Figure 7.10: **Simulation Results of the *Dynamic Network Structure* Experiment.** This figure depicts the simulation results of using a dynamic underlying network structure for all StackExchange.com data sets (top) and all Semantic MediaWiki data sets (bottom). The solid gray line with circles shows the empirical activity. The dashed black line shows the simulated activity from the beginning of one month to its end and the dashed green line is the activity of newly added users at the beginning of one month. These two lines result in the solid black line which shows the corrected simulated activity that is compared to the empirical activity. The plots illustrate that simulation of activity is very accurate with an underlying dynamic network structure for all StackExchange.com networks. The simulations of our Semantic MediaWiki networks turned out to be more accurate compared to the original "static" approach, except for the *PracticalPlants* (j) network, where the sudden increase of empirical activity was not simulated at all.

Table 7.4: **Monthly Structural Changes and Calculated Parameters of the *BeerStackExchange* and *ComplexOperations* Networks.** This table describes changes in the amount of nodes and edges of the graph and all calculated parameters for each month ($\tau_1 - \tau_{12}$). *Activity of New Users* lists the activity of new users added to the network at the beginning of each month and *Diff. to Emp. Activity* the difference of simulated activity and actual empirical acvtivity at the end of the simulation of each month. As more users and edges are added to the networks, this difference varies for each month.

BeerStackExchange	$\tau_0 - \tau_1$	$\tau_1 - \tau_2$	$\tau_2 - \tau_3$	$\tau_3 - \tau_4$	$\tau_4 - \tau_5$	$\tau_5 - \tau_6$	$\tau_6 - \tau_7$	$\tau_7 - \tau_8$	$\tau_8 - \tau_9$	$\tau_9 - \tau_{10}$	$\tau_{10} - \tau_{11}$	$\tau_{11} - \tau_{12}$	τ_{12}
Users	165	198	217	244	271	299	323	341	356	383	408	436	461
Edges	603	677	721	805	851	903	947	973	994	1046	1101	1153	1189
κ_1	18.68	18.99	19.08	19.25	19.31	19.39	19.43	19.46	19.47	19.53	19.58	19.62	19.64
Ratio	19.06	19.22	19.22	19.41	19.46	19.61	19.67	19.62	19.62	19.6	19.68	19.94	–
a_c	2.12	0.73	0.44	0.5	0.35	0.33	0.22	0.14	0.11	0.17	0.23	0.26	0.12
p_{max}	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.0
g	0.01	0.02	0.03	0.02	0.02	0.03	0.05	0.06	0.08	0.05	0.03	0.05	0.03
q	8.71	3.19	1.98	2.27	1.69	1.63	1.14	0.74	0.59	0.92	1.27	1.46	0.71
Activity of New Users	350	71	44	71	38	50	38	26	19	46	50	59	37
Diff. to Emp. Activity	0.00	−0.44	−0.55	−0.81	−0.87	−0.86	−0.53	−0.3	−0.28	−0.28	−0.67	−0.88	−0.23

ComplexOperations	$\tau_0 - \tau_1$	$\tau_1 - \tau_2$	$\tau_2 - \tau_3$	$\tau_3 - \tau_4$	$\tau_4 - \tau_5$	$\tau_5 - \tau_6$	$\tau_6 - \tau_7$	$\tau_7 - \tau_8$	$\tau_8 - \tau_9$	$\tau_9 - \tau_{10}$	$\tau_{10} - \tau_{11}$	$\tau_{11} - \tau_{12}$	τ_{12}
Users	263	264	264	266	268	268	268	268	269	269	272	280	285
Edges	452	452	452	452	452	452	452	452	452	452	452	452	452
κ_1	10.68	10.68	10.68	10.68	10.68	10.68	10.68	10.68	10.68	10.68	10.68	10.68	10.68
Ratio	11.31	10.69	10.88	10.96	10.82	10.68	10.67	10.75	10.63	10.91	10.77	10.55	–
a_c	0.95	0.02	0.02	0.02	0.04	0.01	0.01	0.01	0.01	0.04	0.02	3.15	10.16
p_{max}	0.04	0.04	0.05	0.03	0.04	0.03	0.03	0.02	0.03	0.13	0.04	0.05	0.45
g	0.05	0.92	0.96	0.91	0.7	0.98	0.98	0.83	0.91	0.96	0.88	0.02	0.04
q	8.27	0.42	0.46	0.32	0.51	0.3	0.3	0.18	0.33	1.24	0.43	29.24	96.17
Activity of New Users	249.0	4.0	0.0	4.0	11.0	0.0	0.0	0.0	3.0	0.0	6.0	880.0	21.0
Diff. to Emp. Activity	−0.0	−0.01	−0.34	0.01	0.01	−0.03	−0.35	−0.95	−0.33	−3.11	0.01	−0.03	−448.95

Table 7.4 lists the changes in users, edges and parameters performed on the *BeerStackExchange* and *ComplexOperations* network after the complete simulation of each month. Initially, the simulations of *BeerStackExchange* started with $1,892$ users and $4,000$ edges and after 12 months the network reached a total amount of $2,023$ users and $4,048$ collaboration edges. Hence, the network experienced only small changes during this time. The differences in simulated activity and actual empirical activity range from -0.23 to -0.88.

The *ComplexOperations* network is the smallest one among all our data sets. It started with 263 users and 452 unique links between these users. At the end of simulation, the network reached a total amount of 285 users and the initial amount of collaboration edges, meaning that newly joined users did not interact at all during the observed 12 months. The error of simulation ranges from -0.01 to -448.95.

All results of the other data sets used in this work showed similar behavior: The error of simulation changes as more and more users join the network. We sum up the results for each of the remaining StackExchange.com and Semantic MediaWiki networks in the following paragraphs.

A higher changing of user numbers experienced the *BitcoinStackExchange* data set. Here, the network initially started with 4,567 users and 13,910 collaboration edges which summed up to 6,289 unique users and 17,805 links between them until the end of simulation. The network grew about half the size from the beginning of the simulation over the observed 12 months. The differences between empirical and simulated activity vary between -1.84 and -4.21.

The *ElectronicsStackExchange* network started out with 13,007 users with 83,419 unique interactions between those users. At the end of the simulation, we observed an increase in the number of users by 8,793 to a total of 21,800 users with 120,602 links between them. We experienced a difference between simulated and empirical activity ranging from -2.21 to -12.62 during simulation.

The *PhysicsStackExchange* network initially had 14,334 participants with 85,350 collaboration edges. During the observed time, these numbers changed to 23,608 and 129,065 with 9,274 users that have joined the network within 12 months. Even though the network is one of our larger data sets, we obtained good simulation performance with differences between simulated and empirical activity ranging from -2.79 to -20.34.

The *GamingStackExchange* data set is the largest among all StackExchange.com networks we have used in this work, starting with 24,957 users and 107,786 collaboration edges in the first month of simulation. The network increased to 34,445 users and 131,983 edges over the observed 12 months. The accuracy of simulation changes in each month, with differences in observed empirical and simulated activity ranging from -1.59 to -5.64.

During the observed time, the *BioInformatics* network attracted only 6 new users, lifting the number of users in the first month from 302 to 308 at the end of simulation. These users initially engaged in 310 unique collaborations which summed up to 314 over the 12 months. The differences between

simulated activity resulting from the *Activity Dynamics* framework and real empirical activity range from -1.49 to -33.01.

A bit larger is the *NeuroLex* network, with 845 users and $1,638$ collaboration edges at the beginning of simulation. Over the course of 12 months, 121 new users joined the network which results in a total of 966 users and $1,867$ links between these users at the end of simulation. Depending on each month, the variances between simulated activity and empirical activity range from -6.32 to $-584,39$ for this data set.

The *DotaWiki* network initially started out with $1,892$ users and $4,000$ collaboration edges. Over 12 months, the network managed to acquire 131 new users, which results in a total amount of $2,023$ users and $4,048$ collaboration edges. The accuracy of the activity simulation of this network was average, with errors ranging from -0.11 to -27.83 over all 12 months.

The largest data set among the Semantic MediaWiki networks is the *PracticalPlants* Wiki with only 63 unique users and 75 edges between these users at the beginning of our records. However, it was able to attract $2,156$ new users within the whole observed time span which ultimately results in a total amount of $2,219$ users and 148 collaboration edges. The accuracy of the simulation is varying from -1.23 to -7344.97 for all 12 months.

8 Discussion

This chapter covers our interpretation and discussion of the results covered in Chapter 7. It is split into three sections, corresponding to our three experiments conducted as part of this master's thesis: the *Activity Dynamics Scenarios* explained in Chapter 4, the *Centrality Analysis* explained in Chapter 5 and the introduced *Dynamic Network Structure* explained in Chapter 6.

8.1 Activity Dynamics Scenarios

The six plausible real-world scenarios defined in this work and the results of simulations, as well as the commonalities between the used empirical data sets, allow us to make assumptions about the outcome of real-world events that can take place during a collaboration network's lifetime. Even though the simulated activity resulting from the *Activity Dynamics* framework performs well, we consider the following assumptions only as plausible. More on that can be found in Section 8.4, where we briefly mention the *limitations* of our three experiments. The following paragraphs cover the interpretation and discussion for each of the six scenarios that we have implemented in this work.

Mass Emigration. The outcome of this scenario suggests that when unimportant users leave the network, it is negligible for website owners as activity is not decreasing significantly. However, when highly connected users leave, results showed that activity is influenced, as it drastically decreases. Smaller networks that we used in this scenario seem to better cope with the loss of a high number of important users, whereas large networks seem to struggle to keep acceptable levels of activity whenever high amounts of important

users leave the network. This fact indicates that users in small networks are equally important and connected, while larger networks have only a few highly connected users. One possible explanation for this occurrence is that users with a high degree connect different components of the whole network and removing them also removes the connections between these components. Hence, peer influence is not playing such an important part anymore. This is simultaneously strengthening the effects of the intrinsic activity decay. In this case we suggest that website owners and administrators should find ways to reconnect separate components in order to slow down the decrease of activity whenever a high degree user leaves their networks.

Mass Immigration. The results of this scenario suggest different behavior among the used empirical data sets. One would expect an increase in activity, whenever users are added and connected to existing users. However, this is not the case for the two conducted approaches. Randomly connecting the new users with existing ones brings an increase of activity along, but when we specifically connect them with high degree users, our results showed opposed behavior. One possible explanation for the observed decrease in overall activity is that newly added users start with zero activity. In the beginning of the scenario simulation, the activity goes back and forth between new users and existing high degree users through the mechanism of peer influence. Once the peer influence, that existing high degree users are gaining from the newly added users, reaches a certain level, the intrinsic decay is also increased, and therefore resulting in a decrease of overall activity. To set this observations in context with a real-world scenario, we argue that high degree users are influenced by their peers increasing their activity until a certain threshold is reached where they become over-exhausted and their activity level starts to decrease again. To prevent this occurrence, website owners could implement a recommendation system for newly joined users that preferentially suggests users with a low degree.

Breaking Collaborative Ties. Our results of this scenario suggest similar behavior compared to the *Mass Emigration* scenario. Removing edges between randomly selected users is not significantly affecting the network's activity, whereas the specific selection of high degree users leads to a decrease in overall activity. We explain this by the importance of these users in collaboration networks. According to our model, these users have the most

influence, reaching the largest amount of other users and therefore have the possibility of starting cascades of activity. When removing connections from those users, the whole network might break into different components, making collaborations less likely and less peer influence is impacting existing users within the network. However, even when removing half of all the edges and therefore drastically decreasing the network's overall activity, the level of activity does not reach zero, indicating that resulting small components (or partial networks) still experience activity. As it is hard to predict when and where collaborations between users break, we suggest to monitor the last time of interaction between connected users. Once a defined threshold is reached by two connected users, they could be informed about past activities, eventually bringing them back together.

Establishing New Collaborations. The results of this scenario show different observations for the StackExchange.com networks and the Semantic MediaWiki networks. In general, the introduction of new collaboration edges between existing users increases activity. However the *random* approach does this almost not notably. New connections between high degree users do increase activity more effectively and already at small amounts of added edges (10% of existing edges). Furthermore, when we add more edges, the activity does not relatively increase proportional to that amount as, for example, the *ElectronicsStackExchange* network (Figure 7.4(c) in Section 7.1) reveals. Our empirical Semantic MediaWiki networks show a different behavior, as the *informed* approach performs tremendously better than the *random* approach. These circumstances could be explained by possible over-exhaustion of high degree users already mentioned in the discussion and interpretation of the *Mass Immigration* scenario. As the Semantic MediaWiki networks have a lower average degree and users are equally important and therefore a higher range of users is affected, activity is spreading among more users and therefore preventing the over-exhaustion. Again, website owners could implement recommendation systems that specifically suggests users with low degree in order to prevent the over-exhaustion.

Providing Incentives. This scenario suggests similar behavior for all our empirical data sets. The overall activity of networks will increase as long as 10 posts per month and per incentivized user are injected to the collaboration network. The activity of these users will sooner or later start to spread across the whole network and will inspire other users to start contributing to the

network. However, when incentivized users are not paid and do not inject 10 posts per month anymore, activity will start to fall back to normal levels as other users are not motivated by incentivized users anymore. Based on our results, the enhancement of activity through this approach is advisable for website owners.

Emergence of trolls. We expected the overall activity of networks to decrease in this scenario and our obtained results strengthen this assumption. Smaller networks are more prone to trolls than larger and more active networks. This could be explained by the fact that in larger networks, trolls perish in the large amount of users, whereas in smaller networks all users might notice the troll and are negatively influenced by them. However, when specifically connecting trolls to high degree users in large and active collaboration networks, the activity is also negatively affected and decreases more notably. Whenever one single troll is added, our networks are able to resemble the shape of normal activity, but at lower levels. Five or more *informed* trolls can already bring activity to zero, resulting in a "dead" network as long as no external positive influence is introduced to the network. This is due to the fact that all five trolls are connected to the same important users and negatively affect them via their peer influence in the *informed* approach. Even though trolls do not necessarily harm the network's overall activity, we recommend website owners and administrators to quickly deal with them.

8.2 Centrality Analysis

In this section we interpret the results of the *Centrality Analysis* experiment explained in Chapter 5.

Our results show that more centralized users also end up with higher simulated activity. The centrality of users is strongly depending on the connections to their neighbors and since edges represent the interactions between users, the ones with many edges have more activity as they gain more peer influence. Centralized users have the highest activity weights after simulation, but they still do not reach their empirical activity. We explain this by the intrinsic activity decay being more influential than peer

influence on resulting activity weights of users. However, the centrality of users or the number of connections to their neighbors does not automatically imply the amount of resulting activity. For example, there could be one user with only one single edge to one neighbor, but still half of the network's activity can flow over this single edge as these two users might be especially active. This might explain why some of the least centralized users do not reach their real empirical activity during the simulation process as centrality does play an important role on the activity gained over the peer influence.

All of our used centrality measures show strong correlation between resulting simulated activity, suggesting that future activity of users could be predicted by simply considering the structure of a collaboration network.

Deegre. The degree of users strongly correlates with simulated activity. We expected this behavior due to the fact that the amount of activity gained through the peer influence mechanism is depending on the degree of a user.

Eigenvector Centrality. At the initialization of the simulation, we update users with activity weights depending on empirical activity and the users's eigenvector centrality. The resulting simulated activity *linearly correlates* with the eigenvector centrality of users, suggesting that the non-linear part of the *Activity Dynamics* model is not or only minimally considered. After further investigation, we discovered that peer influence values, calculated with $g_i(a_i, a_j)$ of the *Activity Dynamics* framework, are too close to zero. Hence, peer influence is not affecting the resulting activity in a notably way. We further discuss this occurrence in Section 8.4.

PageRank. The PageRank in an undirected graph matches the degree distribution. Hence, the PageRank also correlates strongly with simulated activity as we have seen for the correlation coefficients calculated for the degree and simulated activity.

The construction of different collaboration networks through the manipulation of the required amount of interactions in order to connect users (denoted by k) did not show any constant improvements compared to the original approach of the *Activity Dynamics* framework. Hence, we suggest that collaboration networks should be constructed as it is already implemented by the *Activity Dynamics* framework.

8.3 Dynamic Network Structure

The results of our third and last experiment, described in Chapter 6, suggest that using an underlying dynamic network structure has advantages compared to the static structure of the original *Activity Dynamics* framework. We significantly decreased the root-mean-square error (RMSE) of the simulated and empirical activity for each data set, except for the *PracticalPlants* network among the Semantic MediaWiki data sets. Here, the RMSE was increased when using the "dynamic" approach. We explain this by the long lasting low values of activity in the first months and the following sudden gain of activity at a later point in time. In the original approach, where all variables and parameters are only calculated once with averages of empirical data over all observed months, this peak is better modeled and included in the ratios. Hence, activity in the months before this peak can already reach higher levels of activity. When we use the "dynamic" approach we simulate low levels of activity more accurately, leaving us with too little activity at the beginning of the month to cover the sudden increase of activity. Even when our calculated ratios better reflect the real-world empirical data, activity weights are too low to move away from a state close to zero activity. However, as long as we are not dealing with activity levels close to zero, our implemented dynamic network structure improves the performance of activity simulations.

In our approach we calculate all variables and the parameter λ/μ for each month. However, the additional run time needed for this calculation is not significantly higher compared to the original approach of the *Activity Dynamics* framework. Hence, we suggest to use our enhanced version for further experiments in the fields of activity dynamics.

8.4 Limitations

The results of the three conducted experiments in this thesis showed that the *Activity Dynamics* framework is capable of simulating activity dynamics with outstanding performance. Hence, we consider our results of the *Activity Dynamics Scenarios* and the *Centrality Analysis* as reasonable. The

parameter calculation and the underlying network structure for the 10 empirical data sets are considered to be static. However, in a real-world online collaboration network, the underlying network structure is continuously changing and ever evolving, whether by newly joining users, leaving users, new interactions between users or by the breaking of connections between them. In addition to that, we only applied one approach (degree of users) for specifically targeting (*informed* approach) users for our six scenarios. Furthermore, we had no possibility to evaluate the results of our scenarios due to the lack of empirical data for comparison with simulated results. For that reason, we included non-manipulated simulated activity in all our scenario plots to show that simulated activity for all data sets only exhibits minor differences to real empirical activity. With this in mind, we argue that the assumptions made in this master's thesis provide good approximations of the impacts of our plausible real-world scenarios on online collaboration networks.

The discovered linear correlation between simulated activity and the eigenvector centrality of users leaves the simulation with the *Activity Dynamics* framework pointless, as the peer influence (the non-linear part of the dynamical system) has no or only a small effect on simulated activity. Hence, we need to find a better way to estimate the ratio λ/μ. More on that can be found in Section 9.2 *Future Work*.

9 Conclusions

The aim of this master's thesis was to provide owners and administrators of online collaboration networks a useful analysis of the various drivers of activity in such networks. Based on the results of the three experiments conducted and applied on 10 different real-world empirical data sets, we gained insights in the possible behavior of users within an online collaboration network. We conducted these three experiments with the *Activity Dynamics* framework introduced by Walk et al. (2015).

The *Activity Dynamics Scenarios* experiment was about the simulation of six plausible real-world scenarios applied to our empirical data sets. To be able to simulate these different events, we adopted and further extended the *Activity Dynamics* framework. Our results suggest complex interdependencies between new users and high degree users (*Mass Immigration* and *Establishing New Collaborations* scenarios), high amounts of randomly removed collaboration edges between users in order to significantly decrease activity (*Breaking Collaborative Ties*) and quickest gains in activity by adding new collaboration edges (*Establishing New Collaborations*) or by providing incentivized users (*Providing Incentives*). Trolls harm the overall activity of the network by negatively influencing other users (*Emergence of Trolls*). We quantified all differences between non-manipulated dynamics and manipulated dynamics in order to provide the reader a detailed overview of the impacts of our scenarios.

The *Centrality Analysis* experiment provided us with insights into the effects of centralized users on resulting simulated activity in the context of our model. For that, we calculated three different centrality measures (degree, eigenvector centrality and PageRank) of each user and compared them with the resulting simulated activity by calculating the Person Correlation Coefficient. The user's degree and PageRank turned out to strongly correlate with

simulated activity. The eigenvector centrality even linearly correlated with simulated activity, suggesting the disregard of the non-linear part within the *Activity Dynamics* framework, leaving it for improvements possibly done in future work.

In the *Dynamic Network Structure* experiment we further extended the original *Activity Dynamics* framework in order to consider the dynamic network structure of collaboration networks. With the implemented improvements, we can now calculate all parameters for model setup and model initialization for each observed month of our empirical data sets. To compare the simulation performance of the original approach and the new dynamic approach, we calculated the root-mean-square error (RMSE) of simulated and empirical activity for both approaches. Our results suggest that activity dynamics can be simulated even more accurately with an underlying dynamic network structure.

The results of these three experiments can be seen as a first step towards new tools and models for website owners to simulate the impact of various events affecting the dynamics of activity in online collaboration networks.

9.1 Contributions

This master's thesis uses the *Activity Dynamics* framework to simulate activity dynamics in 10 different real-world collaboration networks. Further, this work investigates the impact of six plausible real-world scenarios applied on these data sets and uncovers the importance of user centrality for simulation results. To this end, the original version of the *Activity Dynamics* framework is extended, allowing the natural dynamic network structure of collaboration networks to be taken into account.

9.2 Future Work

The conducted experiments leave us with many further possible improvements for the *Activity Dynamics* framework. Besides the simulation of the

Activity Dynamic Scenarios and the *Centrality Analysis* with the introduced *Dynamic Network Structure*, the implementation of different filters (for example an alpha beta filter) could be used to further improve simulation performance. Furthermore, the uncovering of the non-linear part not being considered in most cases during simulation suggests for finding a better way to estimate the ratio λ/μ, for example through the field of *System Identification*. On the other hand, we could further investigate the initialization of the model and its effect on the non-linear part of activity dynamics. To better evaluate the outcome of our *Activity Dynamics Scenarios*, we could seek and prepare further empirical data sets that actually experienced the simulated events.

Bibliography

Abrams, Daniel M and Steven H Strogatz (2003). "Linguistics: Modelling the dynamics of language death." In: *Nature* 424.6951, pp. 900–900 (cit. on p. 22).

Adar, Eytan, Lada Adamic, et al. (2005). "Tracking information epidemics in blogspace." In: *Web intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM international conference on*. IEEE, pp. 207–214 (cit. on p. 12).

Aggarwal, Charu C, Yan Xie, and S Yu Philip (2012). "On Dynamic Link Inference in Heterogeneous Networks." In: *SDM*. SIAM, pp. 415–426 (cit. on p. 12).

Aral, Sinan and Dylan Walker (2012). "Identifying influential and susceptible members of social networks." In: *Science* 337.6092, pp. 337–341 (cit. on p. 24).

Axelrod, Robert (1997). "The dissemination of culture a model with local convergence and global polarization." In: *Journal of conflict resolution* 41.2, pp. 203–226 (cit. on p. 20).

Baronchelli, Andrea et al. (2006). "Sharp transition towards shared vocabularies in multi-agent systems." In: *Journal of Statistical Mechanics: Theory and Experiment* 2006.06, P06014 (cit. on p. 22).

Barrat, Alain, Marc Barthelemy, and Alessandro Vespignani (2008). *Dynamical processes on complex networks*. Cambridge University Press (cit. on p. 9).

Berger, Fredrich Chrisstoph and Patrick van Bommel (1996). "Personalized search support for networked document retrieval using link inference." In: *Database and Expert Systems Applications*. Springer, pp. 802–811 (cit. on p. 12).

Binney, James J et al. (1992). *The theory of critical phenomena: an introduction to the renormalization group*. Oxford University Press, Inc. (cit. on p. 19).

Bonacich, Phillip (2007). "Some unique properties of eigenvector centrality." In: *Social Networks* 29.4, pp. 555–564 (cit. on p. 45).

Brin, Sergey and Lawrence Page (2012). "Reprint of: The anatomy of a large-scale hypertextual web search engine." In: *Computer networks* 56.18, pp. 3825–3833 (cit. on p. 46).

Castellano, Claudio, Santo Fortunato, and Vittorio Loreto (2009). "Statistical physics of social dynamics." In: *Reviews of modern physics* 81.2, p. 591 (cit. on p. 19).

Castellano, Claudio, Matteo Marsili, and Alessandro Vespignani (2000). "Nonequilibrium phase transition in a model for social influence." In: *Physical Review Letters* 85.16, p. 3536 (cit. on p. 20).

Centola, Damon, Juan Carlos Gonzalez-Avella, et al. (2007). "Homophily, cultural drift, and the co-evolution of cultural groups." In: *Journal of Conflict Resolution* 51.6, pp. 905–929 (cit. on p. 21).

Centola, Damon and Michael Macy (2007). "Complex contagions and the weakness of long ties." In: *American Journal of Sociology* 113.3, pp. 702–734 (cit. on p. 11).

Cha, Meeyoung et al. (2010). "Measuring User Influence in Twitter: The Million Follower Fallacy." In: *ICWSM* 10.10-17, p. 30 (cit. on p. 13).

Chakrabarti, Deepayan et al. (2008). "Epidemic thresholds in real networks." In: *ACM Transactions on Information and System Security (TISSEC)* 10.4, p. 1 (cit. on p. 18).

Christakis, Nicholas A and James H Fowler (2008). "The collective dynamics of smoking in a large social network." In: *New England journal of medicine* 358.21, pp. 2249–2258 (cit. on p. 24).

Clifford, Peter and Aidan Sudbury (1973). "A model for spatial conflict." In: *Biometrika* 60.3, pp. 581–588 (cit. on p. 19).

Cosley, Dan et al. (2010). "Sequential Influence Models in Social Networks." In: *ICWSM* 10, p. 26 (cit. on p. 14).

Crystal, David (2000). *Language death*. Ernst Klett Sprachen (cit. on p. 22).

Danescu-Niculescu-Mizil, Cristian et al. (2013). "No country for old members: User lifecycle and linguistic change in online communities." In: *Proceedings of WWW* (cit. on p. 24).

Ferguson, Neil M, Derek AT Cummings, et al. (2005). "Strategies for containing an emerging influenza pandemic in Southeast Asia." In: *Nature* 437.7056, pp. 209–214 (cit. on p. 16).

Ferguson, Neil M, Matt J Keeling, et al. (2003). "Planning for smallpox outbreaks." In: *Nature* 425.6959, pp. 681–685 (cit. on p. 16).

Flache, Andreas and Michael W Macy (2006). "Why more contact may increase cultural polarization." In: *arXiv preprint physics/0604196* (cit. on p. 21).

Flache, Andreas and Michael W Macy (2007). "Local convergence and global diversity: The robustness of cultural homophily." In: *arXiv preprint physics/0701333* (cit. on p. 21).

Galam, Serge and Frans Jacobs (2007). "The role of inflexible minorities in the breaking of democratic opinion dynamics." In: *Physica A: Statistical Mechanics and its Applications* 381, pp. 366–376 (cit. on p. 20).

Ganesh, Ayalvadi, Laurent Massoulié, and Don Towsley (2005). "The effect of network topology on the spread of epidemics." In: *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*. Vol. 2. IEEE, pp. 1455–1466 (cit. on p. 18).

Gomez Rodriguez, Manuel, Jure Leskovec, and Andreas Krause (2010). "Inferring networks of diffusion and influence." In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 1019–1028 (cit. on pp. 11, 12).

Gomez Rodriguez, M et al. (2011). "Uncovering the Temporal Dynamics of Diffusion Networks." In: *28th International Conference on Machine Learning (ICML 2011)*. International Machine Learning Society, pp. 561–568 (cit. on p. 12).

González-Avella, Juan Carlos, Mario G Cosenza, and K Tucci (2005). "Nonequilibrium transition induced by mass media in a model for social influence." In: *Physical Review E* 72.6, p. 065102 (cit. on p. 21).

Goyal, Amit, Francesco Bonchi, and Laks VS Lakshmanan (2010). "Learning influence probabilities in social networks." In: *Proceedings of the third ACM international conference on Web search and data mining*. ACM, pp. 241–250 (cit. on p. 12).

Granovetter, Mark S (1973). "The strength of weak ties." In: *American journal of sociology*, pp. 1360–1380 (cit. on p. 11).

Gross, Jonathan L and Jay Yellen (2005). *Graph theory and its applications*. CRC press (cit. on p. 45).

Holley, Richard A and Thomas M Liggett (1975). "Ergodic theorems for weakly interacting infinite systems and the voter model." In: *The annals of probability*, pp. 643–663 (cit. on p. 19).

huffingtonpost.com (2011). *Why People Unfriend On Facebook*. URL: http://www.huffingtonpost.com/2011/12/19/why-people-unfriend-on-facebook_n_1158326.html/ (visited on 09/23/2015) (cit. on p. 39).

Hufnagel, Lars, Dirk Brockmann, and Theo Geisel (2004). "Forecast and control of epidemics in a globalized world." In: *Proceedings of the National Academy of Sciences of the United States of America* 101.42, pp. 15124–15129 (cit. on p. 16).

Katok, Anatole and Boris Hasselblatt (1997). *Introduction to the modern theory of dynamical systems*. Vol. 54. Cambridge university press (cit. on p. 8).

Kempe, David, Jon Kleinberg, and Éva Tardos (2003). "Maximizing the spread of influence through a social network." In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 137–146 (cit. on p. 12).

Kephart, Jeffrey O, Gregory B Sorkin, et al. (1997). "Fighting computer viruses." In: *Scientific American* 277.5, pp. 56–61 (cit. on p. 17).

Kephart, Jeffrey O and Steve R White (1991). "Directed-graph epidemiological models of computer viruses." In: *Research in Security and Privacy, 1991. Proceedings., 1991 IEEE Computer Society Symposium on*. IEEE, pp. 343–359 (cit. on p. 17).

Kephart, Jeffrey O and Steve R White (1993). "Measuring and modeling computer virus prevalence." In: *Research in Security and Privacy, 1993. Proceedings., 1993 IEEE Computer Society Symposium on*. IEEE, pp. 2–15 (cit. on p. 17).

Kermack, William O and Anderson G McKendrick (1927). "A contribution to the mathematical theory of epidemics." In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. Vol. 115. 772. The Royal Society, pp. 700–721 (cit. on p. 14).

Kermack, WO and AG McKendrick (1932). "Contributions to the Mathematical Theory of Epidemics. II. The Problem of Endemicity." In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, pp. 55–83 (cit. on p. 14).

Kirk, Roger (2007). *Statistics: an introduction*. Cengage Learning (cit. on p. 46).

Klemm, Konstantin et al. (2003). "Global culture: A noise-induced transition in finite systems." In: *Physical Review E* 67.4, p. 045101 (cit. on p. 21).

Lappas, Theodoros et al. (2010). "Finding effectors in social networks." In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 1059–1068 (cit. on p. 13).

Leskovec, Jure, Lada A Adamic, and Bernardo A Huberman (2007). "The dynamics of viral marketing." In: *ACM Transactions on the Web (TWEB)* 1.1, p. 5 (cit. on p. 10).

Lloyd, Alun L and Robert M May (1996). "Spatial heterogeneity in epidemic models." In: *Journal of theoretical biology* 179.1, pp. 1–11 (cit. on p. 15).

Longini, Ira M et al. (2005). "Containing pandemic influenza at the source." In: *Science* 309.5737, pp. 1083–1087 (cit. on p. 16).

Lu, Qing and Lise Getoor (2003). "Link-based classification." In: *ICML*. Vol. 3, pp. 496–503 (cit. on p. 12).

Markus, M Lynne (1987). "Toward a "critical mass" theory of interactive media universal access, interdependence and diffusion." In: *Communication research* 14.5, pp. 491–511 (cit. on p. 7).

May, Robert M and Roy M Anderson (1984). "Spatial heterogeneity and the design of immunization programs." In: *Mathematical Biosciences* 72.1, pp. 83–111 (cit. on p. 15).

Mazzitello, Karina I, Julián Candia, and Víctor Dossetti (2007). "Effects of mass media and cultural drift in a model for social influence." In: *International Journal of Modern Physics C* 18.09, pp. 1475–1482 (cit. on p. 21).

Milgram, Stanley (1967). "The small world problem." In: *Psychology today* 2.1, pp. 60–67 (cit. on p. 6).

Minett, James W and William SY Wang (2008). "Modelling endangered languages: The effects of bilingualism and social structure." In: *Lingua* 118.1, pp. 19–45 (cit. on p. 22).

Mobilia, Mauro (2003). "Does a single zealot affect an infinite group of voters?" In: *Physical Review Letters* 91.2, p. 028701 (cit. on p. 20).

Mobilia, Mauro and Ivan T Georgiev (2005). "Voting and catalytic processes with inhomogeneities." In: *Physical Review E* 71.4, p. 046102 (cit. on p. 20).

Mobilia, Mauro, A Petersen, and Sidney Redner (2007). "On the role of zealotry in the voter model." In: *Journal of Statistical Mechanics: Theory and Experiment* 2007.08, P08029 (cit. on p. 20).

Mooney, Christopher Z (1997). *Monte carlo simulation*. Vol. 116. Sage Publications (cit. on p. 19).

Newman, Mark (2010). *Networks: an introduction*. Oxford University Press (cit. on pp. 9, 15, 45, 46).

Oliver, Pamela, Gerald Marwell, and Ruy Teixeira (1985). "A theory of the critical mass. I. Interdependence, group heterogeneity, and the production of collective action." In: *American journal of Sociology*, pp. 522–556 (cit. on p. 7).

Olson Jr, Mancur (1965). "Logic of collective action: public goods and the theory of groups, The.." In: *Harvard Economic Studies* (cit. on p. 7).

Parravano, A, H Rivera-Ramirez, and MG Cosenza (2007). "Intracultural diversity in a model of social dynamics." In: *Physica A: Statistical Mechanics and its Applications* 379.1, pp. 241–249 (cit. on p. 21).

Pastor-Satorras, Romualdo et al. (2014). "Epidemic processes in complex networks." In: *arXiv preprint arXiv:1408.2701* (cit. on p. 15).

Putnam, Robert D, Robert Leonardi, and Raffaella Y Nanetti (1994). *Making democracy work: Civic traditions in modern Italy*. Princeton university press (cit. on p. 6).

Rvachev, Leonid A and Ira M Longini (1985). "A mathematical model for the global spread of influenza." In: *Mathematical biosciences* 75.1, pp. 3–22 (cit. on p. 15).

Solomon, Jacob and Rick Wash (2014). "Critical Mass of What? Exploring Community Growth in WikiProjects." In: (cit. on p. 8).

Stark, Rodney and William Sims Bainbridge (1980). "Networks of faith: Interpersonal bonds and recruitment to cults and sects." In: *American Journal of Sociology*, pp. 1376–1395 (cit. on p. 6).

Steels, Luc (1995). "A self-organizing spatial vocabulary." In: *Artificial life* 2.3, pp. 319–332 (cit. on p. 21).

Strogatz, Steven H (2001). "Exploring complex networks." In: *Nature* 410.6825, pp. 268–276 (cit. on p. 9).

Strogatz, Steven H (2014). *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Westview press (cit. on p. 9).

Sznajd-Weron, Katarzyna and Jozef Sznajd (2000). "Opinion evolution in closed community." In: *International Journal of Modern Physics C* 11.06, pp. 1157–1165 (cit. on p. 19).

Tang, Jie et al. (2009). "Social influence analysis in large-scale networks." In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 807–816 (cit. on p. 13).

techcrunch.com (2007). *YouTube Launches Revenue Sharing Partners Program, but no Pre-Rolls*. URL: `http://techcrunch.com/2007/05/04/youtube-`

launches-revenue-sharing-partners-program-but-no-pre-rolls/ (visited on 09/23/2015) (cit. on p. 40).

theguardian.com (2013). *Facebook loses millions of users as biggest markets peak*. URL: http://www.theguardian.com/technology/2013/apr/28/facebook-loses-users-biggest-markets/ (visited on 09/23/2015) (cit. on p. 37).

theverge.com (2012). *YouTube opens Partner program to all: every creator in 20 countries can now monetize video*. URL: http://www.theverge.com/2012/4/13/2945243/youtube-partner-program-monetization/ (visited on 09/23/2015) (cit. on p. 40).

theverge.com (2013a). *BlackBerry says BBM messaging app for Android and iOS isn't coming this week*. URL: http://www.theverge.com/2013/9/23/4763994/blackberry-says-bbm-for-android-and-ios-not-coming-this-week/ (visited on 09/23/2015) (cit. on p. 38).

theverge.com (2013b). *How comment trolls can influence your opinions*. URL: http://www.theverge.com/2013/1/11/3865154/comment-trolls-influence-opinions/ (visited on 09/23/2015) (cit. on p. 41).

theverge.com (2015). *Twitter CEO: 'We suck at dealing with abuse'*. URL: http://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the/ (visited on 09/23/2015) (cit. on p. 42).

Van Mieghem, Piet, Jasmina Omic, and Robert Kooij (2009). "Virus spread in networks." In: *Networking, IEEE/ACM Transactions on* 17.1, pp. 1–14 (cit. on p. 18).

Waddington, Jeremy and Colin Whitston (1997). "Why do people join unions in a period of membership decline?" In: *British Journal of Industrial Relations* 35.4, pp. 515–546 (cit. on p. 6).

Walk, Simon et al. (2015). "Activity Dynamics in Collaboration Networks." In: *arXiv preprint arXiv:1505.01634* (cit. on pp. 1, 4, 23, 26, 27, 88).

Wang, William SY and James W Minett (2005). "The invasion of language: emergence, change and death." In: *Trends in ecology & evolution* 20.5, pp. 263–269 (cit. on p. 22).

Wang, Yang et al. (2003). "Epidemic spreading in real networks: An eigenvalue viewpoint." In: *Reliable Distributed Systems, 2003. Proceedings. 22nd International Symposium on*. IEEE, pp. 25–34 (cit. on p. 18).

washingtonpost.com (2015). *How Facebook knows who all your friends are, even better than you do*. URL: http://www.washingtonpost.com/news/the-

intersect/wp/2015/04/02/how-facebook-knows-who-all-your-friends-are-even-better-than-you-do/ (visited on 09/23/2015) (cit. on p. 39).

Watts, Duncan (2007). "Challenging the influentials hypothesis." In: *WOMMA Measuring Word of Mouth* 3.4, pp. 201–211 (cit. on p. 13).

Watts, Duncan J (2002). "A simple model of global cascades on random networks." In: *Proceedings of the National Academy of Sciences of the United States of America* 99.9, pp. 5766–5771 (cit. on p. 11).

Watts, Duncan J and Steven H Strogatz (1998). "Collective dynamics of 'small-world'networks." In: *nature* 393.6684, pp. 440–442 (cit. on p. 11).

Weidlich, Wolfgang (1971). "THE STATISTICAL DESCRIPTION OF POLARIZATION PHENOMENA IN SOCIETY†." In: *British Journal of Mathematical and Statistical Psychology* 24.2, pp. 251–266 (cit. on p. 18).

wired.com (2013). *Instagram User Numbers Down; Updated Terms of Service in Effect This Week*. URL: http://www.wired.com/2013/01/instagram-terms-users/ (visited on 09/23/2015) (cit. on p. 37).

Yorke, James A, Herbert W Hethcote, and Annett Nold (1978). "Dynamics and control of the transmission of gonorrhea." In: *Sexually transmitted diseases* 5.2, pp. 51–56 (cit. on p. 15).