Master's Thesis

Categorization and Analysis of Socialbot Attacks in Online Social Networks

Silvia Mitter, BSc

Graz University of Technology

Knowledge Technologies Institute
Head: Prof. Dr. Stefanie Lindstaedt

Supervisor: Univ.-Doz. Dr. techn. Markus Strohmaier
Advisor: Dipl. Ing. Claudia Wagner

Graz, February 2013

Masterarbeit

Kategorisierung und Analyse von Socialbot Attacken in Online Sozialen Netzwerken

Silvia Mitter, BSc

Technische Universität Graz

Knowledge Technologies Institute
Vorstand: Prof. Dr. Stefanie Lindstaedt

Begutachter: Univ.-Doz. Dr. techn. Markus Strohmaier
Betreuerin: Dipl. Ing. Claudia Wagner

Graz, Februar 2013

Abstract

Online social networks (OSNs) such as Twitter or Facebook are popular and powerful since they allow reaching millions of users online. They are also a welcome target for *socialbot* attacks. Socialbots are autonomous agents in OSNs having their own account, performing certain actions on their own, and maybe mimicking human behavior. Since socialbots could endanger the balance of an OSN's ecosystem, a deep understanding of the nature and potential impact of such attacks is essential.

In this thesis a taxonomy for categorizing socialbot attacks is proposed, based on examples of socialbot attacks conducted on different OSNs. The aim of the taxonomy is to identify similarities and differences of attacks in order to create a better understanding of socialbot attacks and give an overview about possible threats.

Furthermore, this work investigates users which are attacked by socialbots. The users' behavior is analyzed with the aim to find characteristics suitable to identify which users can be successfully attacked. Models are learned for predicting which users are more susceptible to socialbot attacks than others. Results suggest that users which tend to interact with socialbots are in general very active users and communicate with many other users. However, characteristics for users susceptible to attacks on OSNs vary for distinct datasets.

In a next step, the social impact of socialbots is analyzed by measuring if and how the social graph of an OSN can be shaped by specific bot interaction. The findings from this study suggest that socialbots may indeed have a social impact and may have the power to motivate users creating new follow links to others. However, one has to study users' and socialbots' activities carefully over time in order to estimate impact of socialbots. Results also suggest that for a considerable part of newly created links information about the motivation cannot be found in the data available.

This work is relevant for engineers and scientists since it gives an overview about ongoing socialbot attacks and shows how they can be categorized. Furthermore, potential impact of such attacks is investigated in detail to create a better understanding for the problem domain.

Kurzfassung

Soziale Netzwerke im Internet wie Twitter und Facebook haben in den letzten Jahren kontinuierlich an Bedeutung gewonnen. Durch solche Netzwerke können schnell und einfach große Mengen an Benutzern erreicht werden, wodurch auch verstärkt *Socialbots* angelockt werden. Mithilfe solcher automatisierter Agenten können Attacken auf andere Benutzerkonten ausgeführt werden. Um zu verhindern, dass das Gleichgewicht eines sozialen Netzwerkes maßgeblich gestört wird, ist ein tiefes Verständnis für solche Attacken nötig.

Basierend auf bekannten Socialbot Attacken auf unterschiedliche soziale Netzwerke wird eine Taxonomie vorgeschlagen, die Ähnlichkeiten und Unterschiede von Attacken hervorheben und das Verständnis für davon ausgehende Gefahren schärfen soll.

Da Socialbot Attacken ein noch sehr unerforschtes Gebiet darstellen, wird im empirischen Teil dieser Arbeit das Verhalten attackierter Benutzer untersucht. Ziel der Analyse ist es, Eigenschaften und Verhaltensmuster zu identifizieren, die charakteristisch sind für erfolgreich attackierte Benutzer. Diese Charakteristiken werden dann verwendet, um Modelle, für die Vorhersage welche Benutzer anfällig für Socialbot Attacken sein werden und welche nicht, zu lernen. Ergebnisse zeigen, dass Benutzer die eher anfällig für Attacken sind, generell sehr offen und sozial sind, indem sie zum Beispiel mit vielen unterschiedlichen Benutzern kommunizieren. Sowohl die Anzahl der anfälligen Benutzer als auch die Treffsicherheit der Vorhersagen variieren für unterschiedliche untersuchte Datensätze.

In einer weiteren Studie wird betrachtet, welchen sozialen Einfluss Socialbots erreichen können. Es wird untersucht, wie viele neue soziale Links basierend auf Empfehlungen der Socialbots zwischen Benutzern erstellt werden. Ergebnisse zeigen, dass Socialbots durchaus für solche spezifischen Aufgaben geeignet sein können, eine Kausalität aber nur schwer messbar ist.

Diese Arbeit ist nützlich für IngenieureInnen und WissenschaftlerInnen, da sie dazu beiträgt das Verständnis für Socialbot Attacken zu verbessern. Die Taxonomie gibt einen Überblick über ausgeführte Attacken. Im empirischen Teil der Arbeit wird gezeigt, welche Methoden und Kennzahlen verwendet werden können, um Benutzerverhalten und mögliche Auswirkungen von Attacken zu studieren.

Statutory Declaration

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 marked all material which has been quoted either literally or by content from the used sources.

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 entnommene Stellen als solche kenntlich gemacht habe.

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

This work was written at the Knowledge Technologies Institute at the University of Technology in Graz.

First of all I want to thank my supervisor Dr. Markus Strohmaier for the excellent support during my master's thesis. He always took time to discuss questions and problems, especially at the time when we were defining the topic and research questions his ideas were essential for this work. The combination of his valuable input and critical view was very helpful during the whole time.

I also want to express my special thanks to my academic advisor Dipl. Ing. Claudia Wagner. She continuously shared her knowledge and experience in the context of this work and regarding scientific working in general. During several discussions and meetings she gave me most helpful feedback and motivation.

Additionally I want to thank colleagues, friends and the research group I was working with. They always motivated me with feedback, mathematical advice, and gaming nights.

Tim Hwang and colleagues Ian Pierce and Max Nanis made the datasets available for my empirical studies. I want to thank them for their confidence and explanation of the data.

Finally, I want to express my thanks to my family for constantly supporting me through my studies.

Contents

List of Figures

1 Introduction

Online Social Networks (OSNs) have become very popular over the last years. For example, *Armin Wolf* (News Anchorman of Austrian Television) provides his almost 70,000 *Twitter* followers with information about ongoing news and events in the form of short messages. Bands, movies and companies have their own *Facebook* pages to keep fans and customers updated about concerts, new products and other kinds of news. OSNs have become a very powerful communication tool. The downside is that OSNs are not only used for legitimate purposes. They are also used to distribute undesired advertisements, to spread misinformation and propaganda, or to manipulate users.

Recently, several challenges and experiments have been performed on OSNs to study how well automated accounts can embed themselves in the social graph, or how many users are willing to communicate with autonomously acting accounts, mimicking human behavior: Boshmaf u. a. (2011) created more than 100 automated accounts on Facebook designed to befriend complete strangers. The research showed that on average about 35% of the Facebook users accepted such a friend request. The proportion of positive responses even increased up to approximately 80%, if the accounts sent friend requests to users they had common friends with.

Another experiment was performed on Twitter in 2011 by web (2011). Three automated accounts, mimicking humans and human behavior were created to investigate if and how users interact with such accounts. Abusive usage of OSNs in the context of elections have been made public. For example, Thomas u. a. (2012) report that automated accounts were used on Twitter in the context of the Russian parliamentary election. The automated accounts behaved in a way that Twitter keywords, formerly used to organize protests regarding the election, became useless. Other reports claim that a large proportion of followers of the election candidates in the USA[1] and Mexico[2] were automated accounts which should support the candidates.

Several detection mechanisms have been proposed to avoid adversarial campaigns. For example, Gao u. a. (2010) use clustering techniques to identify accounts performing spam campaigns. In Yang u. a. (2011a) measures are introduced to distinct between benign and adversarial user accounts. A lot of research effort was conducted on the creation of powerful bot detection mechanisms. Nevertheless, only little research has focused so

[1] `http://yhoo.it/NPVZ9f`

[2] `http://www.technologyreview.com/news/428286/twitter-mischief-plagues-mexicos-election/`

far on creating a deeper understanding of the nature and impact of actions performed by autonomous agents in OSNs. Therefore this work concentrates on this rather new problem domain. Related terms are introduced below for uniform usage throughout this thesis.

Socialbots: Autonomous agents having their own user accounts in OSNs and performing certain actions on their own. Socialbots do not necessarily pursue a malicious goal.

Targets: User accounts that socialbots try to bond with, e.g. by communication attempts or friend requests and follow behavior.

Socialbot Attacks: Socialbots are considered to attack targets on the OSN if some but not necessarily all of the following characteristics are fulfilled:

1. Socialbots may try to pursue a variety of latent, obscure goals such as to spread information or to influence users.
2. Socialbots may mimic human behavior and/or humans and therefore fake their real identity.
3. Socialbots may exhibit adversarial or malicious behavior.

The characteristics may be seen as *symptoms* for a socialbot attack.

Links: A link describes a social relation between two users if they are friends (bidirectional OSN) or from one user to another one if the one user follows the other one (unidirectional OSN).

The aforementioned research demonstrates that modern security defenses, such as the Facebook Immune System (Stein u. a., 2011), are not prepared for detecting or stopping a large-scale infiltration caused by socialbots. The research community does not have a deep understanding of socialbot attacks on OSNs and their potential impact. Therefore this work presents an exhaustive literature review on socialbot research and proposes a taxonomy of socialbot attacks. The taxonomy helps identifying similarities and differences of socialbot attacks. Further, two empirical studies are performed. The first study aims to investigate behavior of targeted users to predict which users tend to be more susceptible to socialbot attacks than others. The second study concentrates on the potential impact of socialbots to shape the social graph of an OSN by trying to create links between targeted users.

Therefore three research questions are formulated in the next section.

1.1 Research Questions and Contribution

Below, research questions addressed in this work and the contributions of this thesis are summarized.

RQ 1: *How can socialbot attacks on OSNs be categorized?*

Based on extensive studies of ongoing socialbot attacks on OSNs in the last couple of years a taxonomy is developed to categorize socialbot attacks on OSNs. The taxonomy allows to characterize attacks to support experts as well as non-professionals to analyze similarities and differences. The contribution of the proposed taxonomy is to create a better understanding for the broad and constantly growing field of socialbot attacks on OSNs. The utility of the taxonomy is shown by using it to describe investigated socialbot attacks on OSNs from the past.

RQ 2: *To what extent is it predictable whether a user will be susceptible to a socialbot attack or not? Do susceptible users show any specific characteristics which allow to differentiate them from non-susceptible users?*

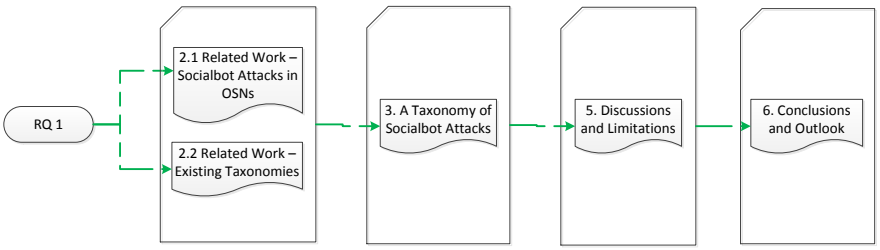
First a study is conducted analyzing to what extent one can predict if a user becomes susceptible during a socialbot attack or not. The contribution of this experiment is to investigate specific characteristics of users, which may help to differentiate between susceptible and non-susceptible users.

RQ 3: *To what extent and how can socialbots manipulate the link creation behavior of OSN users?*

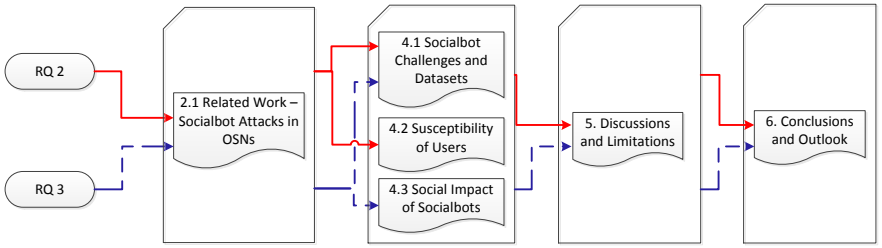
The contribution of this experiment is to identify how successful socialbots can be in creating links between users. It is analyzed to what extent they can be used to shape the social graph. Measures are introduced which allow to assess the success of such link creation attempts by socialbots. By using these measures, the impact of socialbots can be analyzed in more detail as it was possible before and as it was done in previous studies such as Nanis u. a. (2011).

1.2 Organization of this Thesis

In chapter 2 related work regarding socialbot attacks, OSN defense mechanisms and taxonomy creation is reviewed. A taxonomy for socialbot attacks on OSNs is proposed in chapter 3. First, an overview of the taxonomy is given in section 3.1 followed by a description of the methodology in section 3.2. Aspects related to the quality of the taxonomy are presented in section 3.2.1 and the applicability of the taxonomy is shown in section 3.2.2. In chapter 4 empirical studies conducted on two datasets are described. First, the susceptibility of users is investigated and analyzed in section 4.2, then an empirical study regarding the social impact of socialbots is discussed in section 4.3. In chapter 5 limitations of this work are summarized. Chapter 6 discusses conclusions of this work and gives an outlook on possible future work. A visual overview about the relevant chapters along the three research questions is given in Figure 1.1.



(a)Relevant chapters for Research Question 1



(b)Relevant chapters for the empirical Research Questions 2 and 3

Figure 1.1: Research Questions and chapters in which they are addressed.

Parts of this thesis regarding Research Question 2 (see Chapters 4 and 5) were developed in collaboration with two colleagues and have been published in Wagner u. a. (2012). Parts of this thesis regarding Research Question 1 and Research Question 3 (see Chapters 2, 3, 4, 5 and 6) are currently under review at Mitter u. a. (2013). The author of this thesis was not involved in nor did participate in the design, setup or execution of the socialbot challenges described in section 4.1. However, the datasets of these challenges are used for the empirical studies conducted in this work.

2 Related Work

This chapter first reviews different kinds of socialbot attacks on OSNs as well as possible countermeasures and defense mechanisms. Then an overview of literature regarding taxonomy creation is discussed.

2.1 Socialbot Attacks on OSNs

Since socialbot attacks on OSNs are a common threat nowadays, ongoing work related to recent attacks and countermeasures or defense strategies is reviewed in this section.

This work mainly concentrates on research regarding Twitter, therefore this OSN is explained briefly in the next paragraph.

Twitter[1] is a microblogging service where a user can create an account and post several messages which are called *Tweets*. Each Tweet may be built by up to 140 characters. Twitter supports unidirectional links – sources of follow links are called *Followers*, whereas the targets of follow links are called *Friends* (previously the term *Followees* was established). Users can be mentioned in tweets by including a @-sign followed by the username (e. g. *@username*). If one wants to directly address or reply to a user the *@username* has to be written at the beginning of the tweet. Other users' tweets can also be repeated which is called *retweeting*. Keywords are called *hashtags* and can be created by writing *#keyword* and are used to group together entries related to specific topics. Based on the most frequently used hashtags *trending topics* are created.

2.1.1 Attacks and State-of-the-art Technologies

In the last years OSNs arose but these new platforms not only attracted legitimate users. The new and easy way to reach a large number of people also led to an increase in malicious usage of socialbots. A great deal of research effort has been expended in creating countermeasures against spam bot attacks. However, the nature of all kinds of socialbot attacks is still relatively unexplored. Hence recently the research community started to

[1] `https://twitter.com/`

take a closer look on how socialbots can infiltrate social networks, the impact they can have and also how providers of OSNs can prevent their users from socialbot attacks. Different articles regarding the nature and evolution of socialbots are for example discussed in the *social mediator forum*, described in Hwang u. a. (2012). This work reports that socialbots could pursue creditable goals, such as connecting groups of people that are in conflict or try to increase the participation rate in political elections. The downside is that socialbots could also be used for latent and adversarial goals, such as manipulating human users without their knowledge. Hence, the usage of socialbots raises severe ethical questions also in the case of socialbot utilization for research purposes.

In the following, several socialbot attacks performed on different OSNs are reviewed.

Socialbot Attacks on aNobii

Aiello u. a. (2012) reviewed their study conducted on *aNobii* [2] which is an OSN where book reviews can be discussed. An experiment where a socialbot interacts with legitimate users was performed for research purposes. The objective was to test how popular and influential a socialbot with no initial social trust, no profile information and no human behavior mimicking interactions may become. Their findings are alarming. The experiment started *accidentally* – initially the socialbot was used to collect user data to study the nature and dynamics of link creation on aNobii. To this end, the socialbot visited different user profiles in regular time intervals. Then OSN functionality was changed to leave traces when users were visiting other users' profiles. The bot started to leave traces and it aroused interest. Users started to visit its profile and asked questions. After this experiment the socialbot's account was one of the most popular accounts on aNobii. It received about 2,700 messages, as well as 66,000 profile visits and more than 200 followers. The researchers calculated more sophisticated popularity measures such as *PageRank* and *HITS*. Also these measures showed a high popularity rank for the socialbot. Since the bot received such surprising results, a second socialbot was created to perform even more studies. That bot was created in a far more active way. Its objective was to create so called recommender messages to measure if the bot could be used to shape the social graph. Therefore a recommender system was built which was trained with a classifier to create recommendations based on previous information. Then target groups were built where users were split into socialbot following and non following accounts which then were further split into groups with random and advanced recommendation strategies. A fifth group was created, where both users received a recommendation message for each other. Overall 3,000 recommendations were sent, all during one night. Reactions were captured during the next 36 hours, afterwards the bot was removed by the OSN operators. During this phase more than 50% of newly created links were recommended

[2] `http://www.anobii.com/`

by the bot and massive reactions from the users on the socialbot's public wall have been measured. A closer look at the data gave following insights:

- Users following the socialbot were more susceptible to recommendations than others.
- More social links were established based on trained recommendations than on random recommendations.
- Recommendations which both users received showed a much higher success rate than one-sided recommendations.

Although the socialbots did not follow any adversarial objectives results are alarming, since bots also could be used in a more malicious way. Since aNobii is a rather small network not having experienced many large-scale attacks before, the question arises if users are more naive and susceptible to attacks on such a network than elsewhere. To address this question attacks on more popular OSNs are reviewed in the following.

Socialbot Attacks on Twitter

Thomas u. a. (2011) provided insights into several large-scale Twitter attacks. They analyzed more than 1.1 million suspended Twitter accounts which created approximately 80 million Tweets collected through seven months. Suspended accounts were mainly used in spam attacks from which the five largest attacks are discussed in detail. Attackers used URL-shortening services, underground-markets to buy Twitter accounts and spam *affiliation* programs. Affiliation programs offer a commission to anyone who directs users to their websites and the users buy something as a follow-up action. The largest identified attack, *Afraid*, consisted of 124,000 accounts which created more than 14 million tweets each containing at least one shortened URL. They used two different ways to spread affiliation URLs. First they authored tweets directed to users they were not socially linked to, and second they made use of trending topics to spread URLs to a large audience by combining unrelated hashtags with affiliation URLs. This way attackers were able to reach 11.7 million distinct users without the effort of collecting followers first. The campaign lasted six months. A large proportion of accounts performing the attacks were created for the sole purpose to be used in attacks. Furthermore large affiliation campaigns in the context of *Amazon, Clickbank, Speedling and Yuklumdegga* were analyzed in Thomas u. a. (2011).

Further Grier u. a. (2010) investigated spam attacks on Twitter and observed that about 84% of the accounts used for analyzed attacks were *compromised accounts* which means that they originally belonged to benign users but were controlled by adversaries at the time of the malicious behavior. In a follow-up work Thomas u. a. (2012) extensively investigated a large-scale socialbot attack on Twitter conducted at the end of 2011. Attackers used more than 25,000 fraudulent accounts to create over 440,000 tweets attacking the

top hashtags strongly related to the Russian parliamentary elections in December 2011. Tweets combined hashtags used for organizing protests and discussing topics around the election, with unrelated content. This way the attackers introduced a lot of noise into the meaning of targeted hashtags, with the result that they could not be used for their original purpose any further. Existing spam-as-a-service marketplaces can be used to buy malicious accounts on Twitter, email addresses, network proxies and hosts for fraudulent activities. Researchers were able to identify more than 50% of the accounts tweeting to the top 20 hashtags related to the Russian election as spam accounts. Performed attacks were organized as *Sybil Attacks* which are described in more detail later on in this work in section 2.1.2. The attack included compromised hosts all over the world from which more than 39% of the IP addresses were blacklisted for malicious activities before. Mail addresses used to register the accounts showed four distinct patterns regarding naming conventions. After applying those patterns to all available mail.ru addresses about 975,000 accounts all showing the same pattern were uncovered. It was concluded that identified accounts were created and offered in evolving underground-markets. Only 3% of the detected accounts were involved in the election attack. Since Twitter has IP based restrictions in account registration 84% of the socialbots were registered with unique IPs but only 49% of those addresses were used for login later.

In autumn 2011, a research group from California, the *Pacific Social Architecting Corporation* (PacSocial) deployed socialbots on Twitter to study the ability of socialbots to influence the tweet and follow behavior of human users. In the report Nanis u. a. (2011) researchers presented their results where they described that after launching the socialbots, follow link creation between users in the target groups increased by an average of 43%. The PacSocial research team shared resulting data, enabling further analysis regarding the social impact of socialbots, discussed later on in chapter 4.

In *Project Realboy* socialbots with basic behavior and strategies were implemented a few years ago, as described in Coburn u. Marra (2008). Greg Marra, the creator of this project, described his strategy in Hwang u. a. (2012). He referred to physical robots which are often based on the *sense-think-act* paradigm. The paradigm describes that at the beginning a robot has to observe its surrounding to get an impression of obstacles and possible ways. Second it thinks about a strategy to pursue its goal. In a third step it decides which strategy is best and acts on it. These steps are repeated until the final goal is reached. In project realboy, this paradigm was applied to socialbot behavior on Twitter. The first step was all about collecting information. In this phase new bots were launched, and data was crawled. Then, socialbots searched for common interests of other users by applying clustering algorithms to identify communities by states. In the acting phase this information was used to target users. The socialbots tweeted about previously identified topics. Since this project was not created to deal with advanced natural language processing the socialbots simply copied the content of different tweets around identified topics and did not mark them as retweets. With this strategy the project realboy bots were able to receive an overall follow-back rate of approximately

33%. This project again showed the potential of socialbots, even if not much effort or costs were invested since rather simple strategies were used.

In the beginning of 2011, a socialbot challenge was organized by the Web Ecology Project web (2011) on Twitter. The two-week challenge took place from January to February 2011. Three teams of socialbot developers participated. Three main bots were released to target a group of 500 users chosen by their common interest in *cats*. Socialbots tried to mimic humans by creating corresponding profile information as well as mimic human behavior in their tweets. The bots were able to interact with 202 out of the 500 targets during the challenge. Now, since Tweets are restricted to a maximum of 140 characters one could argue that it is easier to successfully embed socialbots in the social graph on Twitter than on other OSNs. However, Boshmaf u. a. (2011) showed that this is also possible on other OSNs.

Socialbot Attacks on Facebook

In 2011, researchers launched a Socialbot Network (SbN) consisting of 102 socialbots and one botmaster on Facebook, as described in Boshmaf u. a. (2011). First the researchers ran their bots for a duration of 8 weeks where they sent 8,570 friend requests to randomly chosen targets from which 3,055 were accepted which is an average acceptance rate of 35.7%. Additionally researchers were interested in how the friend-acceptance rate changes when users have common friends with the socialbots. Therefore the socialbots were run for another 6 weeks where requests were sent to friends of already successfully targeted users. The acceptance rate increased to an average of 59.1% and up to 80% when at least 11 mutual friends could be measured. This indicates that the trust factor increases with the number of common friends. Depending on the privacy settings of user accounts collecting data on Facebook is not limited to just directly befriended accounts but rather information from indirectly befriended accounts (the *extended neighborhood*) can be retrieved. The SbN was able to collect information of 1,088,840 profiles including the extended neighborhood.

Gao u. a. (2010) also inspected malicious activities on OSNs. They identified spam activities carried out by more than 57,000 accounts on Facebook. A closer look at those accounts showed that approximately 97% of them were compromised accounts. The study concentrated on large scale socialbot attacks using Facebook's wall messages where about 70% of the identified messages included URLs to phishing sites.

Socialbot Attacks on amazon.com

Jindal u. Liu (2008) directed their research towards *Opinion Spam* which is a common term for spam in product reviews, blogs and forum posts. A dataset crawled from *ama-*

zon.com in June 2006 with more than 5.8 million reviews for 6.7 million products was investigated. The research uncovered fake reviews for products and brands. Not many insights about how the attacks were accomplished were given in Jindal u. Liu (2008).

In the recent past the research community showed a large interest in searching for reliable defense mechanisms and detection methods for all kinds of socialbots. Detecting and defending against socialbot attacks is an *adversarial learning problem* which means that there is a constant struggle between attackers performing a new kind of attack and defenders constantly reacting and improving defense mechanisms to fend off attacks. This is described in more detail later on in this section by the example of the *Facebook Immune System*. A deep understanding of ongoing developments regarding attack detection and defense mechanisms seems essential to create insights for socialbot attacks.

2.1.2 Socialbot Detection and Countermeasures

To create a better understanding of socialbot attacks and gain more insights detection and defense mechanisms are discussed in this section. Many different attempts regarding socialbot detection were made in the past. Starting from manual inspection and identifying possible indicators for socialbot behavior, to investigating different machine-learning approaches using content based, graph based or profile based features with the purpose to train models for automated detection, to applying specific defense mechanisms against sybil attacks or proposing ranking systems of OSN users. In this section a selection of the most salient approaches is presented.

Determining a Groundtruth

One major challenge in getting started with socialbot detection in OSNs is retrieving a dataset with legitimate and adversarial accounts as a groundtruth. Different approaches were used previously. Lee u. a. (2010) and Lee u. a. (2011) described how they run passive accounts called *honeypots* to attract malicious users. The term honeypot is inspired by the common use in network security where a honeypot usually is a computer or a service gauging an attacker to be something or someone else. In the context of OSNs honeypots are accounts used to attract adversarial users and possible attackers. By running a large number of honeypot accounts over seven months on Twitter Lee u. a. (2011) were able to collect a huge amount of followers. Based on the used honeypot strategy no obvious reason was given for benign users to follow those honeypots therefore researchers considered all attracted followers to be bots. Tyagi u. G.Aghila (2011) also described a honeypot-approach for identifying botnets.

Several researchers used blacklisted URLs to reveal socialbots posting a large amount of such URLs as for example shown in Grier u. a. (2010); Gao u. a. (2010); Yang u. a.

(2011a). Other researchers invested much time in manually identifying socialbots and malicious users. Research from Chu u. a. (2010) was based on a Twitter dataset consisting of more than 500,000 users. The large dataset was created by using randomly chosen users as starting points for a Dept-First Search (DFS) algorithm as well as crawling the public timeline for additional user data. Then, a training and a test set were created by manual inspection each consisting of 1,000 human users, bots and *cyborgs* which are defined as semi-automated human assisted bots. Gianvecchio u. a. (2011) crawled a large dataset from *Yahoo!* chat. They manually labeled users from collected data as human, not clearly observable and bots. Other research studies, such as Ghosh u. a. (2012); Thomas u. a. (2011) relied on Twitters suspension mechanism by simply assuming suspended accounts were malicious accounts. Although no description of Twitter's suspension algorithm is publicly available Twitter published *Best Practices and Rules*[3]. Violating those rules can lead to account suspension.

Most of the above mentioned approaches are either expensive in resources and/or introduce a bias.

Receiving datasets to serve as groundtruth directly from OSN providers is not common, since they would have to deal with privacy issues. Also anonymized data still can reveal too much information. Renren [4], the largest OSN in China, made an exception and provided a dataset to researchers for improving their bot defense mechanism. Results from this analysis are described in this thesis in section 2.1.2 .

Machine Learning Approaches

Chu u. a. (2010) collected Twitter profiles to build an automated classification system, to distinguish between human users, bots and cyborgs (brief description of how they built the groundtruth in section 2.1.2). They first calculated an entropy based measure regarding the regularity of a user's tweeting behavior. The entropy indicates whether there is a high or a low instability factor in a user's tweets. The research described that a high entropy indicates a high proportion of uncertainty therefore a random tweeting behavior or a high regularity. A low entropy shows a small proportion of uncertainty and a medium entropy measure indicates a complex therefore human behavior. Second, they used a Bayes classifier highly established in email spam detection to automatically classify tweet content. Different features were calculated per user sample, such as URL ratio, the source from which tweets were created, ratio of friends and followers, and whether accounts were verified or not. As in the research reported, those values seemed well-suited for distinction between bots and humans. Results for a combined classifier showed a high accuracy for distinction between humans and socialbots.

[3] `https://support.twitter.com/articles/18311-the-twitter-rules`
[4] `http://renren-inc.com/en/`

Despite the fact that the results were quite impressive many of the identified features do not seem to be very robust. One can easily change the tweeting time, also the URL ratio can be evaded by simply tweeting many pseudo tweets without links. However, the follower to friends ratio is more elaborate to evade, since a large-scale network would be needed to ensure a high follower number. This underlines the process of *adversarial learning* – as the detection mechanisms improve, the attack strategies improve.

A similar approach regarding chat bot detection was described in Gianvecchio u. a. (2011). They also used entropy based features in combination with a Bayesian classifier to detect chat bots and their results also showed high accuracy even for detecting more sophisticated chat bots. In the past mainly keyword based filtering approaches or puzzle solving approaches were used.

Yang u. a. (2011a) conducted an empirical study on how to detect spammers by defining features and then ranked those features by their robustness against changes in attack strategies. From a pool of 24 features, they labeled four features to have a high robustness:

- *Account Age*: An attack has to be long-term planned or fraudulent accounts have to be bought from underground-markets to change account age. This usually is expensive in time, resources or money.
- *Betweenness Centrality*: A well known measure from graph theory which describes the position of a vertex (user) in a graph. For every vertex pair shortest distance paths are calculated, the more often a specific vertex is part of this path, the higher its betweenness centrality is. Since spammers often tend to befriend randomly chosen or socially unrelated accounts spammers tend to have a high betweenness centrality.
- *Clustering Coefficient*: This is also a graph based measure which describes how related the neighborhood of a vertex in the graph is. Results suggested that socialbots are likely to have a rather small clustering coefficient based on the same reason as they tend to have a high betweenness centrality: they usually follow a high number of unrelated accounts.
- *Followings to Median Neighbors' Followers*: For this measure the two-step neighborhood of a user is considered. It divides the number of users one is following by the median number of those users' followers. This measure ensured that a few accounts with high popularity did not compensate for a lot of unpopular followees (Yang u. a., 2011a). Since bots often follow a large number of users, independent of their position in the social graph, or even target users with just a few followers on purpose, since they may follow back more easily than popular accounts, this measure seems very robust.

Even though these features seem partially robust, they do not seem to be totally stable. For instance, if a socialbot was be able to actively influence the social graph between other users, it could attempt to change values regarding betweenness centrality and

clustering coefficient. Studies from Nanis u. a. (2011) revealed that socialbots may be able to successfully shape the social graph, as discussed also later on in chapter 4.

A cluster-based detection approach was used by Gao u. a. (2010) to detect spam campaigns on Facebook. The research analyzed messages containing URLs by clustering messages with same text patterns or URLs referring to the same websites. Fingerprints were calculated for the descriptive part of a message by using MD5 hash values. If a certain amount of fingerprints between two messages matched these messages were clustered together. This way messages from different authors with a high probability to belong to the same pool of accounts could have been identified. The process of creating and comparing fingerprints was described in more detail in a previous work Zhou u. a. (2003). For the clustering process every message represented a vertex connected to all similar vertices. Gao u. a. (2010) reported that this could theoretically yield in high computational costs but they experienced an acceptable computation time on the provided data since the graph was sparse. In a next step clusters were split in benign clusters and adversarial clusters. This was done by calculating from how many distinct accounts the messages were authored. In email spam detection different IP addresses sending spam are count, this method was simply mapped to Facebook by using account IDs instead of IP addresses. The second indicator used was within which timeframe messages were created. Since socialbots usually expect a rather short service life, they have to maximize their outcome and therefore they are likely to create a large amount of messages in a rather short time period.

User Account Ranking Systems

Another approach to fight spam accounts is to introduce ranking systems. Different ranking algorithms have been proposed. One was to punish users for socially connecting to malicious accounts (Ghosh u. a., 2012). Another one to analyze relationships and also punish users for supporting malicious accounts (Yang u. a., 2012). Another idea was to create a framework based on established ranking measures as the *HITS* algorithm to detect malicious users (Bosma u. a., 2012). This study was conducted on *Hyves*[5], a large Dutch social networking platform.

An excerpt of established and the most salient new importance ranking algorithms is explained briefly:

HITS: This algorithm was developed by Jon Kleinberg in Kleinberg (1999) as a predecessor to the *Page Rank*. One distinguishes between *Authorities* and *Hubs*. Authorities are websites of important content whereas hubs point to authorities they recommend. Each page has a *hub weight* which is proportional to the sum over all *authority weights* it links to, and an authority weight, which is proportional to the sum over hub weights

[5] `http://www.hyves.nl/`

of pages which link to it. This way a page with influential content receives many links and a high authority weight and pages with important links to other pages receive a high hub weight.

PageRank: Google Founders Sergei Brin and Larry Page introduced the PageRank in Page u. a. (1999). It expands the HITS algorithm by introducing a *rank source* which basically is a decay factor to avoid closed loops emerging from sites that link only to each other. Also the *random surfer* is introduced which takes into account that users may jump to completely unrelated websites instead of following links.

Follower Rank: Users are ranked based on their follow count.

Retweeted Rank: Users are ranked based on how often their tweets are retweeted.

Ghosh u. a. (2012) showed that benign users are open to support spammers – they reported that a vast majority of spam supporters were benign users and introduced the term *social capitalists* for users willing to follow back almost everyone. Analyzing the top 100 spam followers revealed that they had a follow-back rate of 80%. Conducting an extra experiment where a fake account was created to follow those social capitalists, showed that the account was able to position itself under the top users according to the PageRank in only three days, by simply following those accounts. Closer investigation of the social capitalists showed that a large portion of them were very popular accounts, according to the follower rank, retweeted rank and PageRank. Inspired by these findings an algorithm to punish spam supporters was created, the *Collusion Rank* (Ghosh u. a., 2012). The algorithm was based on a slightly adopted PageRank. Instead of equally distributed initial values malicious accounts started with a negative value. The second adoption was that a user's score was calculated by its friend's counts instead of its followers based on the idea that one should not be punished for his followers. Applying the collusion rank to the user set on Twitter showed a better performance than the PageRank. 94% of spam accounts were ranked in the *last* 10% whereas with the PageRank approximately 40% were ranked under the *top* 20%.

Similar results were shown in Yang u. a. (2012), where a high proportion of spam followers were highly ranked users. They also differentiated between spam and benign accounts and inspected relations between spammers and between benign users and spammers. A two step algorithm to punish spam accounts following users was proposed which can be used to punish accounts supporting malicious actions. To enable identifying malicious accounts a second algorithm was introduced. The algorithm used identified malicious accounts as seed users and spread scores through the graph. The research showed that combining the two algorithms enabled detection, monitoring and early-stage-ranking of suspicious accounts.

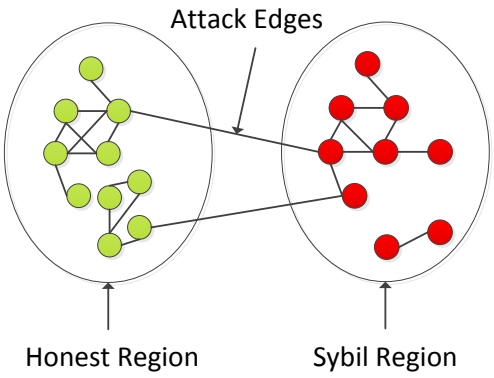


Figure 2.1: Social graph subdivided in honest and sybil region, inspired by Yu (2011).

Defense against Sybil Attacks

So far approaches to detect attacks by using feature extraction for machine learning algorithms and ranking based systems were presented. Defending against *sybil* or *botnet attacks* concentrates on the social graph. In 1973 Flora Rheta Schreiber wrote a book about a woman, called *Sybil*, who had 16 different personalities. Douceur (2002) introduced the term *sybil attacks* according to this book by mapping different personalities from the original character *Sybil* to multiple accounts hold and controlled by one attacker. Hence, sybil attacks are attacks performed by multiple accounts in distributed systems. Sybil accounts are a threat to OSNs because they can conduct large-scale attacks if they are well nested into the social graph. Sybil defenses in trusted networks are based on a trusted authority to bind online identities to real identities. Douceur (2002) reported that if no trusted authority is used defending against sybil attacks is not possible in distributed systems since attackers have the possibility to create more than one account which enables sybil attacks in the first place. Yu (2011) provided a tutorial and survey on how to defend against sybil attacks in social networks.

Yu (2011) gave key insights into sybil defense by leveraging social networks. As shown in Figure 2.1 a distributed system is believed to consist of two types of participants – honest and sybil participants. It can be modeled as a social graph $\mathcal{G} = (V, E)$ consisting of vertices V and edges E. The graph consists of honest and sybil vertices and can be divided into a subgraph of honest vertices and edges between them, a subgraph of sybil vertices and edges between them as well as connecting *attack edges* between honest and sybil vertices (see Figure 2.1).

Sybil and Large-Scale Attacks

Yang u. a. (2011b) used a dataset provided by Renren[6]. Based on a groundtruth of

[6] `http://renren-inc.com/en/`

1,000 sybil identities and 1,000 honest identities, four features used for detecting sybil identities were introduced:

- *Invitation Frequency*: Measuring how many friend requests an account sends within a given time period showed that sybil accounts tend to have a higher frequency than benign ones.
- *Outgoing Requests Accepted*: If friend requests are directed to known users the acceptance rate is likely to be higher than for trying to befriend complete strangers. The resulting acceptance rate compared to the number of sent requests was considerably higher for honest users than for sybil users.
- *Incoming Requests Accepted*: Reverse results were presented for the proportion of accepted requests out of received requests where sybil users showed a higher acceptance ratio.
- *Clustering Coefficient*: as briefly described in 2.1.2 the clustering coefficient indicates how closely related a user's neighborhood is. They showed that honest users tend to have a higher clustering coefficient since they have less but better connected friends.

Features were used to train a classifier for distinguishing between benign and adversarial user accounts which was applied by Renren providers to their OSN. The classifier helped detecting 560,000 sybil accounts.

In a second step, Yang u. a. (2011b) investigated how reliable state-of-the-art sybil defense algorithms are in OSNs which are untrusted networks. Established algorithms such as SybilGuard (Yu u. a., 2008), SybilLimit (Yu u. a., 2010) or SybilInfer (Danezis u. Mittal, 2009) are built on two core assumptions:

1. Sybil accounts tend to build clusters with other sybil accounts by connecting to them to avoid detection.
2. Only a limited amount of attack edges between sybil and honest nodes exist based on a lack of interest in befriending strangers.

Yang u. a. (2011b) analyzed Renren regarding those two core assumptions of established algorithms. The research reported that the assumptions cannot be verified by the OSN. Identified sybil accounts were not heavily connected to each other. Furthermore the found connections did not seem to be built on purpose since they were established at some time but not near account creation. The largest identified sybil cluster showed that many links to honest accounts were established. Hence, the author's main conclusion was that none of the two core assumptions could be made for Renren.

Opposed to these findings for Renren, in Cao u. a. (2012) an algorithm called *SybilRank* was introduced as a sybil defense mechanism on *Tuenti*[7] a large Spanish OSN. The algorithm is built on the two core assumptions mentioned above and the research

[7] `http://www.tuenti.com`

reported that it could be successfully applied to defend against sybil attacks. SybilRank can be applied to undirected networks, since it is based on properties and assumptions that are proven to hold for undirected graphs. Honest regions and sybil regions have to be well connected each but almost no connections should exist between those regions. SybilRank propagates trust through a graph. Some benign user must be known to serve as seed values in the honest region. Starting at these vertices trust is spread through the graph, mainly in the honest region by calculating the probabilities that a *random walk* ends at a specific vertex. The algorithm has to be stopped after $O(\log n)$ steps for best performance.

Vertices are then ranked by the calculated trust values, where vertices with lower trust values are more likely to belong to sybil accounts than others. SybilRank outperforms other state-of-the-art sybil defense mechanisms in following points as reported in Cao u. a. (2012):

- *Computational Expenses*: SybilGuard (Yu u. a., 2008) and SybilLimit (Yu u. a., 2010) use a large number of random walks resulting in high computational expenses. SybilInfer (Danezis u. Mittal, 2009) does not guarantee any upper limit on false rates and also has high computational expenses. SybilRank shows fewer computational expenses.
- *Supporting multiple honest regions*: A common weakness of sybil detection algorithms is that they cannot deal with multiple regions of honest users, because starting with more seed vertices in different regions leads to higher computational expenses. SybilRank's computational expenses are not based on the number of seed vertices, therefore this limitation can be overcome.
- *Unbound Seed Selection*: Attackers could try to bring their sybil vertices close to the seed vertices to increase the probability that they earn high trust. SybilRank does overcome this by not assigning lots of trust to closely positioned vertices. Nevertheless, SybilRank also cannot deal with the situation that sybil vertices earn higher trust than honest vertices which would lead to unreliable results.
- *Accuracy*: Measuring the *Receiver Operating Characteristic* (ROC) on the given dataset SybilRank also outperforms other algorithms, such as SybilGuard (Yu u. a., 2008), SybilLimit (Yu u. a., 2010) and SybilInfer (Danezis u. Mittal, 2009).

Applying SybilRank to an artificial network as well as to Facebook confirmed the observations mentioned above. SybilRank was implemented with Hadoop[8] (a MapReduce framework) and it was also tested regarding its scalability. The prototype additionally was applied to an Amazon cluster. Computational costs increased almost linearly and also the largest graph could have been processed within a proper time (160M nodes, processed in 33 hours).

[8] `http://hadoop.apache.org/`

The evaluation showed that the two main goals of SybilRank could have been reached on Tuenti. One possible conclusion comparing it with Yang u. a. (2011b) is that the behavior of sybil attacks may be biased towards the OSNs they are deployed to.

Boshmaf u. a. (2012) discussed challenges of defending against Socialbot Networks (SbNs) after researchers were performing a socialbot attack on Facebook, described in section 2.1.1. The research analyzed OSN vulnerabilities and defense strategies for avoiding or limiting success of sybil attacks. To this end, Boshmaf u. a. (2012) identified the following challenges regarding web automation identity binding and complexity of security:

- *Reverse Turing Test*: The goal of a reverse turing test is to let a machine decide whether a user is human or not. One way to do this is to use *CAPTCHAs*. Unfortunately CAPTCHAs are not a big challenge any longer since many services offering solving CAPTCHAs are available. The research discusses the possibility of using social information instead but also consider that this may be solved by social engineering technologies.
- *Limit Crawls*: Adversaries can reconstruct a social graph, by exploiting unlimited and unrestricted user information. Possible account or IP limitations can be overcome by sybil networks since they are run from different accounts and often run on different hosting machines. The challenge is to restrict sybil crawls without limiting usability for users.
- *Detect Abusive Usage*: Usually requests sent via http can be distinguished from requests sent over an API. However, several web automation techniques allow mimicking http requests, even looking like they are sent from a web browser.
- *Online-Offline Binding*: The original way of defending against sybil attacks includes trusted authorities. Since OSNs like Twitter and Facebook are untrusted networks such an identity mapping is not supported. Boshmaf u. a. (2012) proposed open identity management involving the government by using open identity technologies. The research reported that trust frameworks would be conceivable but come with their own challenges to solve.
- *Security Settings with better Usability*: The best technical security system is worthless if it is not used properly. A major step in fighting sybil attacks on OSNs is by making users aware of the risks to interact with strangers. Developing defense mechanisms regarding sybil attacks relies on the fact that attackers are not able to build many links to benign users. Due to the fact that this assumption does not necessarily hold (Boshmaf u. a., 2011) it is even more important to provide clear privacy setting options to users to ease decision making.

Boshmaf u. a. (2012) further discussed two approaches for countermeasures against adversaries. However, all of the approaches discussed in this section showed some weaknesses.

OSN Security Defenses

Every successful OSN is built on trust between users. An increasing number of socialbots and socialbot networks could endanger the balance of an OSN's eco system. Platform owners are therefore interested in defending their networks against malicious bot attacks. So far many approaches for socialbot detection have been discussed. Lots of different features for classification approaches were identified in several research studies, passive attraction approaches were discussed, as well as adapting sybil defense mechanisms. One may think that combining the knowledge and findings from the research community should be more than enough to successfully prevent OSNs from socialbot attacks. Despite the wide range of detection and defense mechanisms proposed by the research community, the problem of socialbot attacks and malware is still and maybe more than ever present on OSNs. Is it because platform providers are afraid of annoying their users by mistakenly suspending benign accounts, as suspected in Ghosh u. a. (2011)? Or is it because fighting the adversaries is a complex problem as described in Stein u. a. (2011)? Can a perfect detection mechanism even exist when adversaries are continuously improving their techniques? To answer the question how OSN providers try to improve security of their OSNs the Facebook Immune System (FIS) is discussed in detail as an example.

Stein u. a. (2011) gave insights into the structure and functionality of the defense system FIS. The general goal of FIS is to detect attacks against the social graph as soon as possible to fend them off. Defending the social graph is an *adversarial learning problem*. Attackers are interested in keeping their attacks undetected for a maximum amount of time whereas defenders try to minimize detection time. The latters' goal is to maximize attackers' costs and minimize outcome. An adversarial cycle arises (shown in Figure 2.2).

Three main threats to the social graph were identified in Stein u. a. (2011): First of all, accounts originally created from and used by benign users can be taken over by attackers. Those accounts are especially hard to detect since benign users may start to behave in a malicious way out of the sudden. In the past, features like IP addresses and geolocations were used to detect such accounts, by measuring changes and abnormalities. Adversaries responded by creating botnets using compromised hosts to show different geolocations and IP addresses for avoiding detection. The most effective method to recognize hacked accounts still are user reports telling about them. A second threat are accounts created with the sole purpose to perform malicious activities. Such accounts increase their value the longer they exist. It usually takes some time until accounts can successfully embed into the social graph, hence it is essential that adversarial accounts can be detected as early as possible to avoid nesting into the graph and maximizing costs for attackers. Another threat are users, generally showing benign behavior, but taking also part in undesired actions. Stein u. a. (2011) cite supporting chain letters as example, since they

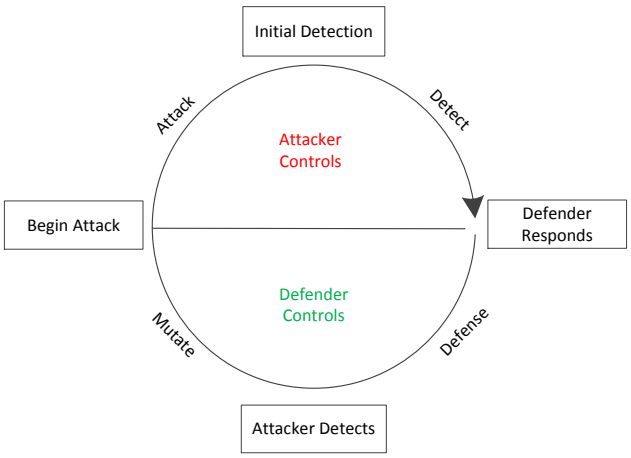


Figure 2.2: The adversarial cycle inspired by Stein u. a. (2011). The diagram shows the constant struggle between attacks and defense against them. First an attack takes place and stays undetected for some time. Right after detection the system has to learn about the attack before a proper reaction can be chosen. In this stage the adversary does not know the attack has been discovered. Finally the attack can successfully be stopped by the defender. At some point the attacker knows that the attack has been detected and starts to change it. This circle repeats from the beginning, and as defense mechanisms improve also attacks become more sophisticated.

can spread around rather quickly. As latest developments showed also services are arising inviting users to set undesired actions against money.

The defense system is designed and organized to be an online learning system. Even if major system updates have to be made services should be updated without taking the system offline. An extra component is developed which ensures new features can be developed and tested in an easy way. A policy engine creates responses to actions and is validated by arising differences between a test group and a validation group. Since the system is built upon an online learning strategy observations are passed to the online services as new features so that the system can learn from them without going offline. The main design components of the defense system and how they are structured were reported in Stein u. a. (2011) in detail.

According to the explanation of the FIS algorithm advantages of the system are scalability and the design to learn and stay online without the need of any restarts. Even though this system seems to be well designed and structured in a modern way, Boshmaf u. a. (2011) reported that when they performed their socialbot attack against Facebook users, only 20% of their socialbots were detected and mainly because of users which marked the accounts as spam. That leads to the assumption that the FIS was not directed to sophisticated socialbot attacks so far. It also indicates that the applied socialbot attack was new to the adversarial circle, which has not been learned by OSN defenders so far.

2.2 Existing Taxonomies

To address the first research question *How can socialbot attacks on OSNs be categorized?* existing literature regarding taxonomy creation and existing taxonomies in the field of *Network and Computer System Attacks* are reviewed below.

Amoroso (1994) discussed threats and attacks to computer systems. The research reported that attack taxonomies could be useful under specific circumstances and summarized three requirements for an attack taxonomy:

- *Completeness*: Taxonomy categories should be exhaustive and since attacks are often complex, unstructured and system dependent, empirical examples often are the strongest way to ensure completeness.
- *Appropriateness*: A reasonable tradeoff between applicability of the taxonomy for one specific system and general but unspecific applicability should be chosen.
- *Internal vs. External Threats*: A differentiation between threats from the inside and the outside of a system should be made to distinguish between different types of attacks.

This is an early work and was often cited in other taxonomy proposals.

Also other research works creating taxonomies in the field of computer attacks identified several properties and requirements for taxonomy creation (Lindqvist u. Jonsson, 1997; Krsul, 1998; Bishop, 1999). Lough (2001) and Hansman u. Hunt (2005) summarized those criterias in their works. Hansman u. Hunt (2005) proposed a taxonomy of network and computer attacks using four dimensions to describe such attacks: One dimension describing how one can reach targets, one dimension defining who is the target, one dimension describing potential vulnerabilities, and one dimension categorizing side effects caused by the attack.

Paulauskas u. Garsva (2006) proposed a *computer system attack* taxonomy that used 14 dimensions to classify attacks. Figure 2.3 shows the identified dimensions and categories.

A taxonomy of *web attacks* was proposed by Alvarez u. Petrovic (2003). The taxonomy was applied to known examples to ensure the quality of the taxonomy. This taxonomy did not use dimensions, but lifecycle steps. Every attack covered an entire lifecycle consisting of nine steps: An attack starts with an *entry point* exploiting a system *vulnerability*. The attack is a threat to a system *service* by completing specific *actions*. The *length* of arguments for HTPP requests and *HTTP elements* are identified as the next lifecycle stages. The *target* is identified as next stage, followed by the *scope* as the effect of the attack and *privileges* describing results after successfully performing the attack.

Quinn u. Bederson (2011) created a taxonomy for *human computation systems* based on empirical examples. This work serves as a reference since it mainly built the taxonomy

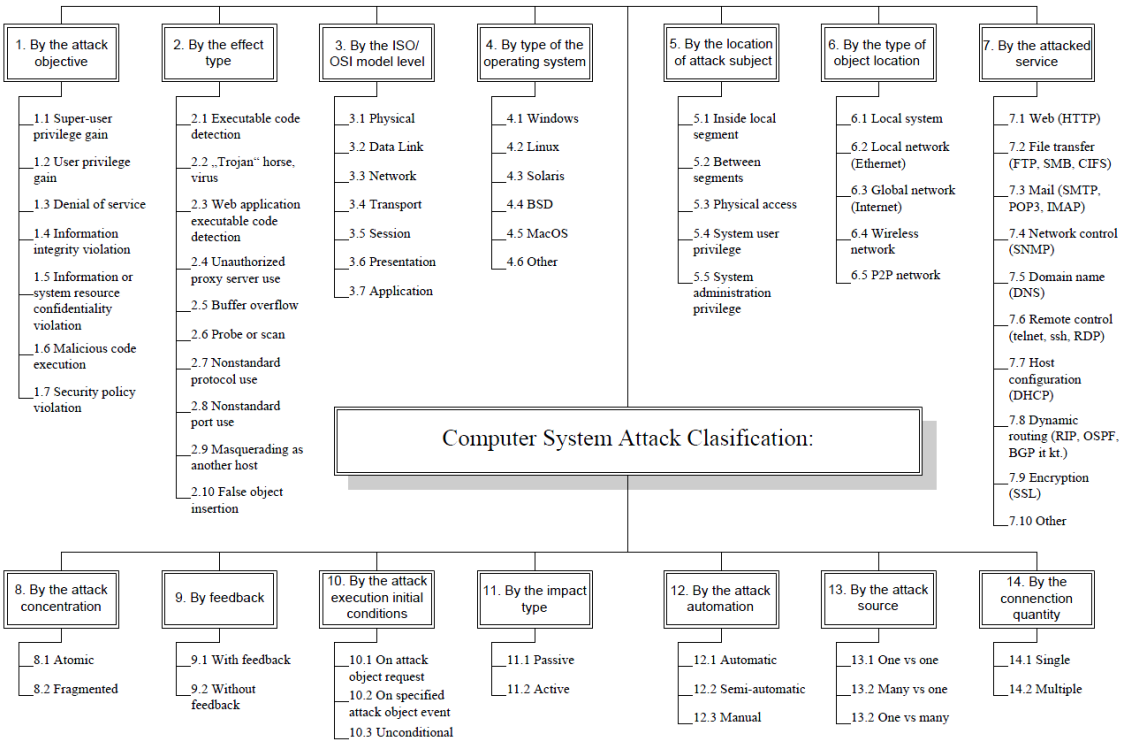


Figure 2.3: Computer system attack classification, from Paulauskas u. Garsva (2006).

on empirical examples in the field of human computation, also using the concept of dimensions for the taxonomy.

Zhao (2003) described conditions under which humans interact with each other as the concept of *copresence*. Two dimensions were identified: the *Mode* of being with others determined by physical conditions and the *Sense* of being with others determined by feelings. One category describing the mode of being with others is the *Virtual Telecopresence*. This means two humans can reach each other digitally, with one actually being present as person and the other one only digitally. *Socialbots* served as an example and were divided into two groups –*instrumental* and *communicative* ones. Instrumental socialbots were described to be used for simple, routine tasks, whereas communicative ones to interact with humans on an emotional basis.

3 Research Question 1 – A Taxonomy of Socialbot Attacks

As described in chapter 2, a lot of research has been done in the field of socialbot detection and defense mechanisms against socialbot attacks in Online Social Networks (OSNs) in the recent past. However, the number and variety of socialbot attacks are growing constantly. According to *Research Question 1 – How can socialbot attacks on OSNs be categorized?* a taxonomy is proposed with the purpose to categorize socialbot attacks on OSNs. The aim of this taxonomy is to give an overview of the variety of socialbot attacks to create a better understanding for this rather new problem domain and related research field. By categorizing socialbot attacks, similarities and differences can be identified. The taxonomy may, for example, serve as a basis to create an understanding of socialbot attacks for smaller OSNs that did not have to deal with such attacks so far.

The taxonomy proposes an exhaustive categorization system and allows to categorize attacks along several dimensions. It rather concentrates on the problem domain than on current state-of-the-art technologies or services.

This taxonomy allows categorizing socialbot attacks along different dimensions in order to describe and compare different attacks. In the following sections the proposed taxonomy is presented in detail, followed by a description of the developing process and a taxonomic description of recent socialbot attacks.

3.1 Overview

The proposed taxonomy allows to categorize socialbot attacks on OSNs. Attacks conducted outside the OSN, or directed against targets outside the OSN are not covered by the taxonomy. A tree-structured hierarchical categorization is not useful for this kind of taxonomy, since it would result in a tree where subcategories would have several parents. Several former taxonomy proposals in related fields also use different dimensions for categorization (e. g. Paulauskas u. Garsva (2006); Alvarez u. Petrovic (2003); Igure u. Williams (2008)). The following dimensions were found to be useful for describing and categorizing socialbot attacks: *Targets*, *Account Types*, *Vulnerabilities*, *Attack Methods*, and *Results*. Every dimension is built as a tree, where leafs represent the categories of

a dimension. Dimensions differ in width and depth, but the level of abstraction within every dimension is consistent. The taxonomy focuses on *Internal Threats*, that means attacks within an OSN since *External Threats* are extensively covered by existing taxonomies regarding general system and web attacks, such as Paulauskas u. Garsva (2006); Igure u. Williams (2008); CVE (2012). Dimensions include theoretical categories for external threats for the sake of completeness, but those categories are out of the scope of this work.

Since the main focus of ongoing research is based on empirical studies and experiments on Twitter or Facebook, the categorization system is biased towards those OSNs. Nevertheless, the taxonomy is believed to be general enough to cover attacks on other OSNs since it concentrates on the general problem domain rather than on OSN-specific problems. The categorization system considers automated socialbot attacks therefore human-based attacks may also be (partially) categorizable by the system although they are not within the focus of the taxonomy.

In the following, taxonomy's dimensions are explained in detail.

3.1.1 Characterizing Targets

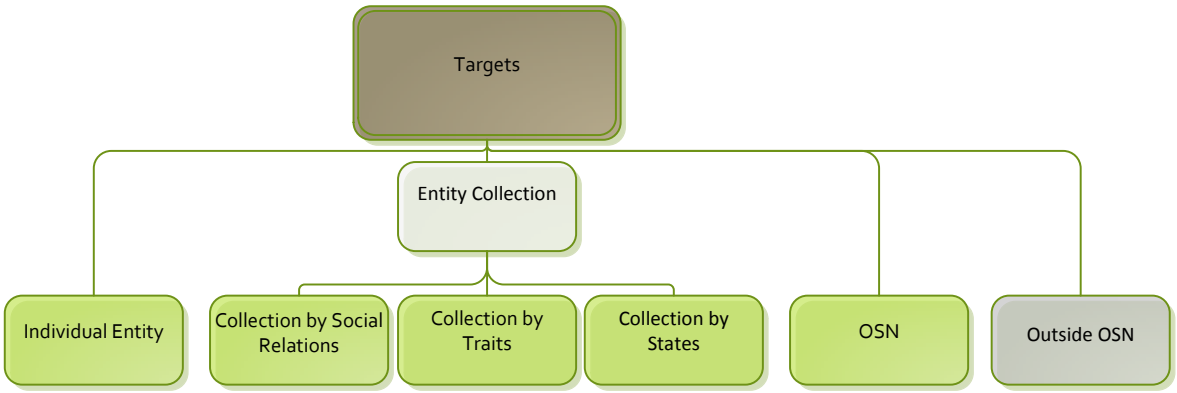


Figure 3.1: Dimension Targets

If a socialbot attack takes place one or several targets are involved. That means for each socialbot attack, one can identify *who* or *what* is attacked, i. e. who or what is the *target* of the attack. Note that who or what is attacked in the first place is not necessarily the same as who or what is harmed in the end (see results dimension). Who receives spam messages? Who is asked to befriend some stranger? What personal site or wall is flooded? Which hashtag is attacked by spam messages? As described in Bishop (1999), categorization should be based on technical details rather than on social causes or effects to avoid speculations.

As shown in Figure 3.1, an individual entity, a collection of entities, or the OSN itself can be the target of an attack. An entity is defined as an item which has its own space within the OSN. Examples of entities can be users who have their own user page, events having their own event page, or locations which have their own location page within the OSN. Which kind of entities are available depends on the specific OSNs. For example on Twitter, user accounts or hashtags can be entities which may be attacked by socialbots.

In Boshmaf u. a. (2011) a Facebook attack from 2011 is described. The attack started by targeting individual user accounts, selected by a random number generator. They attacked a large number of individual entities, which were not chosen by any relations.

A collection of entities can be formed by social relations, by traits (rather static properties of the entities such as the age or the sex of a user entity) or by states (rather dynamic properties such as the user mood or interest of a user entity, or the meaning of a hashtag which can change over time).

Boshmaf u. a. (2011) also investigated how targets react if they have common friends with an attacking stranger, and therefore attacked user accounts related by friendship which serves as example for entity collections by social relations. Organizers of a socialbot competition on Twitter in 2011 web (2011) decided to target a user group, chosen by common states, i. e. all targets where cat lovers and talked about cats. Although the OSN itself could theoretically also be targeted, no example is available. Targets outside the OSN are not in the scope of this work. Defining who is attacked and splitting entity groups by how they are related to each other may help understanding the nature of the attack and also detecting potential other targets.

3.1.2 Characterizing Vulnerabilities

Attacks usually require certain vulnerabilities (see Figure 3.2) which they can exploit in order to facilitate the attack. This dimension describes *what* can be exploited to perform an attack. It differentiates between OSN-specific vulnerabilities and general system vulnerabilities which are out of the scope of this work and have been analyzed and categorized in previous work, e. g. CVE (2012). Vulnerabilities described in this taxonomy focus on specific OSN functionalities and are split into system- or user-caused vulnerabilities.

System vulnerabilities partially emerge from the tradeoff between providing unrestricted, uncensored platform and security. They often enable an attack in the first place. System vulnerabilities can be categorized as follows. Unprotected *Entity Information* can enable potential attackers to retrieve valuable information about a user's relations and activities. Boshmaf u. a. (2012) describe the threat when the social graph of an OSN can be crawled by adversaries, which is exploited in several attacks (e. g. in web (2011); Nanis u. a. (2011); Boshmaf u. a. (2011)). The more information available to friends, or friends of

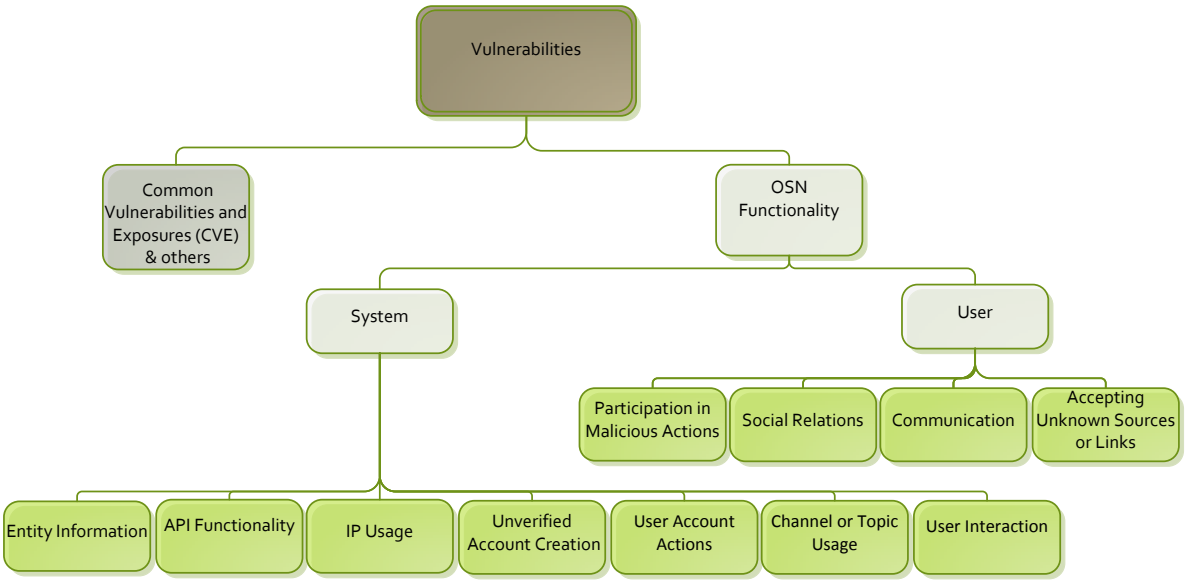


Figure 3.2: Dimension Vulnerabilities

friends, or even the public, the more freedom can be reached but at a higher level of risk.

Comprehensive *API Functionality* can extensively increase the usage of an OSN, since third-party applications can be developed and offered to users and the OSN can be embedded in other websites or systems. The downside is, that no or insufficient API restrictions regarding functionality and the number of API calls ease performing automated attacks, as e. g. shown in Thomas u. a. (2011); web (2011); Nanis u. a. (2011). For instance, the Facebook's Graph API was exploited in an attack in 2011 (Boshmaf u. a., 2011).

Unrestricted *IP Usage* allows adversaries to use one IP address to conduct attacks, so they do not have to manage distributed attacks. OSNs could run tests against online available IP blacklists as e. g. reported in Thomas u. a. (2012) a large scale attack was conducted where about 39% of identified IPs could be found on IP blacklists. On the other hand, IP restrictions can be problematic, since standard users do not use static IP addresses. OSNs would risk to offend legitimate users when blocking or restricting IP addresses, nevertheless missing or inefficient IP restrictions make OSNs vulnerable for attacks.

Vulnerabilities also arise by *Unverified Account Creation* due to missing usage of *Trusted Authorities* as known from the cryptography sector. Usually one only has to provide an email address to create a user account on an OSN. The missing use of trusted authorities means that no control system for verifying accounts is given, meaning everyone can impersonate whoever they want. This vulnerability was exploited by a majority of the

investigated socialbot attacks when creating accounts for their attacks (e. g. Thomas u. a. (2011); web (2011); Nanis u. a. (2011); Boshmaf u. a. (2011); Aiello u. a. (2012)).

If *User Account Actions* are unrestricted, potential attackers can cause a lot of damage in very short time since they can perform (nearly) unlimited actions. Several countermeasures regarding unrestricted access of users are commonly used, such as the *Completely Automated Public Turing test to tell Computers and Humans Apart* (CAPTCHAs). However, several techniques exist to circumvent CAPTCHAs and, as also described in Boshmaf u. a. (2011), ineffective CAPTCHAs are a major vulnerability of OSNs. Restrictions could also help fighting against account hacking attempts. As described in the article about *security holes regarding Twitter accounts*[1] login attempts should be limited per IP and per user account to avoid brute force attacks for guessing passwords.

Unrestricted *Channel or Topic Usage* presents a vulnerability, since everyone can use communication and broadcasting channels and topics such as e. g. hashtag streams in an unrestricted manner. For example, this is the case when one does not have to belong to a specific community in order to use the community's communication channel or topic. Potential attackers can make communication channels and topics useless by adding noise, as shown in the attack against hashtags tied to the Russian election in 2011 (Thomas u. a., 2012).

Unrestricted *User Interactions* allow users to interact with each other regardless of their social relations or other information and restrictions as exploited in several attacks (e. g. web (2011); Nanis u. a. (2011); Boshmaf u. a. (2011)). Independent of how an attacked user reacts, if the system prohibited such communication in the first place, an attack could not take place at all. Nevertheless, it would also limit the freedom of the OSN.

User Vulnerabilities arise since OSNs are especially driven by users and how they behave. They are the most difficult-to-control vulnerability for platform providers, since the providers do not have direct influence on users' behavior.

As reported in Stein u. a. (2011), benign users intentionally *participating in malicious actions*, e. g. repeating a post of a specific advertisement motivated by a chance to win something, are a threat to an OSN's defense system. This way advertisement can spread around quickly and a large number of users can be reached. As described in dimension 3.1.4, services are available where users can sell their accounts for single activities.

Accepting friendship requests from unknown sources to create new *social relations* is another user vulnerability that can be exploited. Several studies show that socialbots are indeed able to create a large amount of relations with unknown users (e. g. web (2011); Nanis u. a. (2011); Boshmaf u. a. (2011)). Another interesting fact described in Boshmaf u. a. (2011), is that it is more likely for users to accept friend requests from

[1] `http://cnet.co/SYycoG`

strangers if they share mutual friends. Socialbots can obviously exploit this for their own good.

Users communicating with unknown users in OSNs can conduce to the success of socialbot attacks (web, 2011; Nanis u. a., 2011). Untrusted *communication* can take place in a variety of ways, for example by using wall postings or private messages. Depending on the concrete communication channel, other users may be influenced by the communication, since communication between users usually implies that they know each other.

When users are tricked to participate in scam actions such as clicking on links or accepting content from unknown sources (such as allowing access for third-party applications) they *accept unknown sources or links*. Grier u. a. (2010) refers to users following affiliation attacks by clicking on URLs.

3.1.3 Characterizing Attack Methods

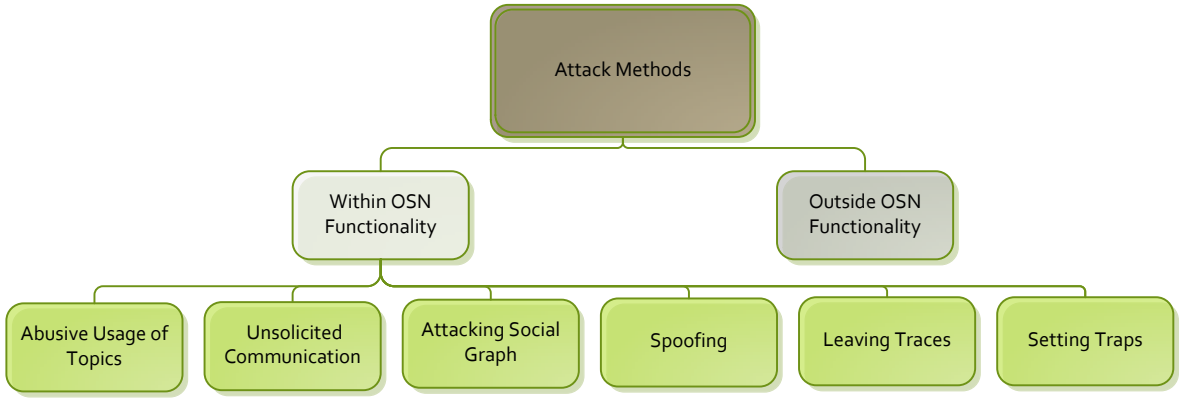


Figure 3.3: Dimension Attack Methods

Different socialbot attacks adopt different attack methods (see Figure 3.3). An attack method describes *how* an attack is applied. Again, the taxonomy concentrates on attack methods within the OSN functionality, and offers one collective category for attacks outside the OSN, as for example a phishing attack via mail to retrieve user credentials for the OSN.

Abusive usage of topics is a common attack method, e. g. described in Jindal u. Liu (2008). Several platforms offer topic specific communication where topics can e. g. be keywords such as hashtags in Twitter or discussion threads on Stack Overflow. Topics often can be used to reach a large audience. Changing the meaning of a topic to a specific new one by combining the topic with different specific context is called *hijacking* topics. Destroying or changing the meaning of a specific topic without assigning a new specific meaning is called *censoring* a topic. Creating faked ratings or reviews for a poll or a

discussion is also called hijacking. Some researchers refer to it also as *Opinion Spam* as described for an attack in Jindal u. Liu (2008). *Clickjacking* in combination with abusive topic usage is combining a topic with links, trying to attract users to click on them. Simple spam attacks are just embedding links in a benign context within a topic. *Affiliation attacks* are used to make a user buy something on a specific website and in return the attackers receive commission. The attacks are conducted by using topics to spread the malicious context. Attacks using hijacking and censoring of topics are for example inspected in Thomas u. a. (2011).

Affiliation attacks either use hijacking or censoring methods or *unsolicited communication* methods in order to perform an attack, by e. g. sending personalized messages to users containing URLs. Unsolicited communication can also be used by e. g. sending directly addressed private messages or mentioning of users in wall posts, as used by several attacks (e. g. in Thomas u. a. (2011); Grier u. a. (2010); web (2011); Nanis u. a. (2011); Gao u. a. (2010)).

Socialbots can *attack the social graph* by trying to befriend targets. OSNs can support only bidirectional friendship relations where the target has to accept a friend request or unidirectional friendship relations where a socialbot may establish a follow relation to the target and may try to make the target follow him back. It is a common attack method which can be found in several socialbot attack investigations, e. g. in web (2011); Nanis u. a. (2011); Boshmaf u. a. (2011).

Spoofing describes the process of disguising ones identity. The problem of impersonated accounts on Twitter is also discussed in an article describing this phenomenon by giving several examples[2].

If a system provides the possibility to *leave traces* on other users' profiles, this feature can be used to create an implicit attack by arousing a user's interest, as for example described in Aiello u. a. (2012).

Traps can be *set* to attract specific users such as benign accounts interested in specific topics, or to uncover malicious accounts wherefore for example *Honeypots* can be used. Honeypots are known from the security sector, where they are used to send adversaries down the wrong track to distract them from real possible system vulnerabilities. On OSNs honeypots are usually passive accounts not trying to befriend someone and acting in a way that benign users usually do not have any reason to actively befriend them. This way a majority of the attracted users may be identified as socialbots, as described in Lee u. a. (2011).

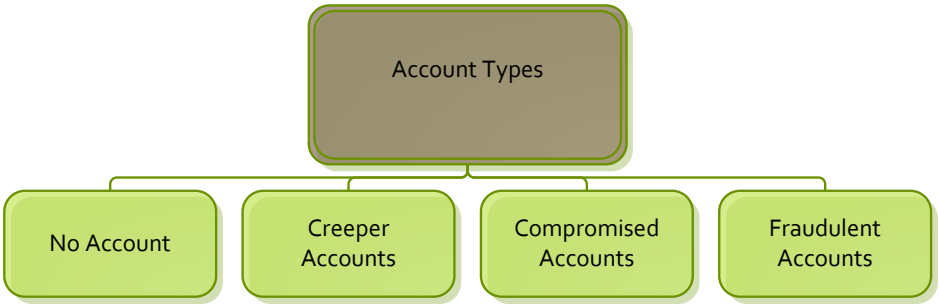


Figure 3.4: Dimension Account Types

3.1.4 Characterizing Account Types

Socialbots require some sort of account type in order to be able to perform an attack. Four different account types are identified (see Figure 3.4).

If OSNs allow user interaction for users that have *no account* then attacks can be applied from within the system, without a specific account. Other attacks conducted without an account from outside the OSN are not in the scope of this work.

Creepers Accounts belong to usually benign users, willing to partially participate in malicious behavior. Stein u. a. (2011) describe creepers as benign users which are using their accounts in an undesired way, as they spread for example chain letters. For some OSNs services offering users to temporarily sell their accounts for single posts are emerging. Examples are *pay4tweet*[3] or *Pay with a Tweet*[4], where attackers can buy single tweets from benign users and use them to advertise their products. This way they reach at least all the followers of the user as audience which they could hardly reach otherwise. The price for one tweet is often based on the account's position in the social graph. Accounts from creepers are hard to detect since they belong to benign users but may show some similarities with compromised accounts since they may start tweeting seemingly random or inappropriate content.

If adversaries illegally start to control accounts from legitimate users such accounts become *compromised accounts* as e. g. described in Stein u. a. (2011). Detecting such accounts can be especially challenging, since a benign user suddenly starts abusive behavior. Such accounts may be well nested in the social graph and therefore reach a large audience which can cause extensive damage. On the other hand, the assumption seems natural that users will report their accounts to be taken over by someone else as soon as they find out. For a majority of compromised accounts at least the active ones this means that attackers can place very effective attacks but only within a short timeframe.

[2] `http://wrightresult.com/2011/12/help-impersonated-twitter/`
[3] `http://www.pay4tweet.com/`
[4] `http://www.paywithatweet.com`

In the recent past some incidents where Twitter accounts were compromised became public[5], Facebook has to deal with stolen user accounts[6] and also *pinterest*[7] (a social network with an online pinboard) has to deal with compromised accounts[8]. Grier u. a. (2010) reported that about 84% of investigated accounts involved in spam attacks were *compromised accounts* and Gao u. a. (2010) observed more than 97% of accounts in spam campaigns were compromised.

Accounts created with the only purpose to be used in a malicious way are commonly called *fraudulent accounts*. Depending on the targeted OSN different ways of retrieving fraudulent accounts are possible. Accounts can be created by the adversary itself or be purchased at so called *spam-as-a-service* marketplaces as e. g. reported in Thomas u. a. (2011). Fraudulent Accounts vary in their appearance strategy from obviously showing that they are fraudulent over giving no profile information away at all to mimicking legitimate accounts for instance by providing an account picture and a biography. While obviously fraudulent accounts and accounts with no or little profile information risk an early detection their creation costs can be minimized whereas advanced accounts could reach much higher acceptance rate within the social graph of an OSN usually also leading to higher costs. Websites offering bulk accounts for OSNs for little money can be easily found online[9]. Fraudulent accounts may be created and used alone or in a bunch. Examples for the use of fraudulent accounts can be found e. g. in Thomas u. a. (2011, 2012); web (2011); Nanis u. a. (2011); Boshmaf u. a. (2011); Aiello u. a. (2012).

3.1.5 Characterizing Results

Finally different socialbot attacks lead to different results (see Figure 3.5), which are defined as the observable outcome of the attack. They are split into active and passive results, depending on whether an active change in the OSN is achieved or not. First active results are described in more detail.

A *changed social graph* is for example reached if socialbots were able to nest themselves in the existing social graph. Other changes, such as newly created or removed social links between users may also result from socialbot attacks but are quite hard to measure since lots of effects outside the OSN could also play a major role in modifying the social graph, as shown later on in the second empirical study in section 4.3.

Socialbot attacks often result in *modified communication channels or topics* (e. g. Jindal u. Liu (2008); web (2011); Nanis u. a. (2011); Aiello u. a. (2012); Thomas u. a. (2011)).

[5] http://bgr.com/2012/11/08/twitter-accounts-hacked-2012/

[6] http://www.pcworld.com/article/247370/ramnit_zeus_hybrid_compromises_45_000_facebook_accounts_what_you_should_know.html

[7] http://pinterest.com/

[8] http://tnw.co/UGdQE1

[9] https://buyaccs.com/en/

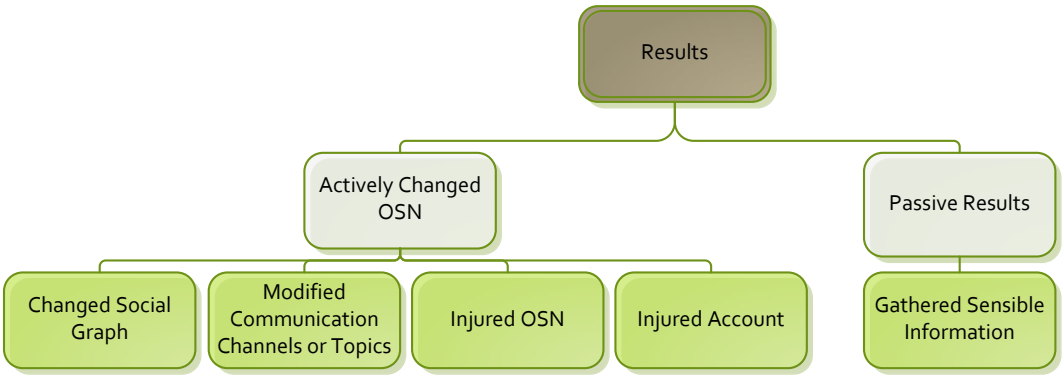


Figure 3.5: Dimension Results

OSNs provide different ways to create discussions or group entries around a specific topic which can be changed in their usability or meaning by attackers.

An *injured system* can be *jammed* or *hacked*. A jammed system has problems to handle the load of requests it is getting, and can be evidenced by long system response time or no system response. A hacked system is at least partially in control of the attacker.

Accounts can be *injured* in different ways, e. g. reputation can be damaged (which would be hard to measure) or accounts may be blocked by the OSN as a consequence of a socialbot attack.

If a socialbot has access to protected personal information of a successfully attacked user this would be an example for *gathering sensible information*.

3.2 Methodology

The taxonomy was created as an iterative process, by executing the following steps:

1. First, existing research in the domain of socialbot attacks in OSNs was inspected, to identify similarities and differences between socialbot attacks as basis to form dimensions and categories.
2. Next, the categorization system was critically inspected regarding the *Six problems of categorization* introduced in Mervis u. Rosch (1981) as a first step to improve quality of the taxonomy.
3. Then the quality of the taxonomy was evaluated by considering properties and requirements for categorization systems, based on previous work (Lough, 2001; Bishop, 1999; Krsul, 1998; Lindqvist u. Jonsson, 1997; Amoroso, 1994; Howard, 1998).

4. The taxonomy was applied to some socialbot attacks to ensure applicability. Example categorization of attacks and references to attacks per dimension and category can be found in section 3.2.2.
5. Finally, the taxonomy proposal was discussed with other researchers.

3.2.1 Quality of Taxonomies

The quality of taxonomies can be assessed in two steps: First, the taxonomy is investigated regarding Mervis' Six problems of categorization then properties and requirements of taxonomies are discussed.

Six problems of categorization

In Mervis u. Rosch (1981) following six problems regarding categorization of objects are discussed.

1. *Arbitrariness of categories*: Categories can be chosen rather arbitrary or in a more natural way. If chosen more naturally it is usually easier to categorize objects and distinct between objects in a useful manner. Mervis u. Rosch (1981) summarized properties which appear to be more suitable to describe natural categories.
2. *Equivalence of category members*: Although, members of a category should be almost equally representative for that category equivalence of members is hard to reach (Amoroso, 1994). Features identified in Mervis u. Rosch (1981) for measuring equivalency of category members are listed below.
    - *Speed of processing* measures the time needed for categorizing.
    - *Order and probability of exemplar production* which shows that the more a member of a category is mentioned the more representative it is for this category.
    - *Natural language terms* are used to indicate the representativeness of objects.
    - *Asymmetry in similarity ratings* relates to known similarity measures, e. g. the *Kullback Leibler Distance* which is an asymmetric measure. Mervis u. Rosch (1981) report a kind of a hierarchy of category members. Members which are little representative for a category tend to be more similar to highly representative members and the other way around - representative members are less similar to less representative members.
    - The *Learning and Development* process describes that highly representative members should be categorized first to improve accuracy of categorizing less representative members.
3. *Determinacy of membership* addresses the problem of boundaries. Boundaries of categories as well as boundaries of a category's items should be well-defined.

4. The *Nature of abstraction* implies that necessary information has to be separated from irrelevant information and meta information has to be extracted from gained knowledge to find a good level of abstraction.
5. *Decomposability of categories into elements* implies that despite categorization is usually based on decomposition some properties should be included as a holistic property.
6. The *Nature of attributes* describes the tradeoff between what is considered to be an attribute and what should be a category.

Properties and Requirements

Based on properties and requirements for taxonomies identified in former work (Lough, 2001; Bishop, 1999; Krsul, 1998; Lindqvist u. Jonsson, 1997; Amoroso, 1994; Howard, 1998) the following requirements can be considered to be important for this taxonomy proposal.

- *Unambiguous and Well-Defined* (Bishop, 1999; Howard, 1998; Lindqvist u. Jonsson, 1997; Krsul, 1998): By extracting common terms from existing literature and investigated socialbot attacks categories are described in a clear, well-defined and unambiguous way.
- *Repeatable and Objective* (Howard, 1998; Krsul, 1998): A deterministic categorization process should lead to the same results for one attack independent of the categorizing person. Socialbot attacks and the taxonomy were discussed in different stages of the creation process with colleagues from the research team to ensure the taxonomy is understandable and attacks would be categorized in the same way.
- *Exhaustive* (Howard, 1998; Lindqvist u. Jonsson, 1997): By using information about conducted socialbot attacks as a survey it is possible to create a rather exhaustive taxonomy. Amoroso (1994) claims that empirical examples are proposed as an indicator for completeness.
- *Useful*(Howard, 1998; Lindqvist u. Jonsson, 1997): The taxonomy is created in a way to be useful to experts as well as users with fewer domain knowledge by giving an overview and referring to existing attacks in detail.
- *Based on Technical Details* (Bishop, 1999): Only technical details are valid concepts. Social causes should not be a valid concept since this would lead to speculations about objectives. The proposed taxonomy follows this principle, e. g. if a Twitter hashtag is flooded with noise the hashtag is the targeted entity not the user audience possibly reading the tweets and therefore may indirectly be influenced by the attack.
- *Similar but Multiple Categorization*: Similar attacks are categorized in a similar way. Mutual exclusiveness is not claimed for this taxonomy, as proposed in other

taxonomies (e. g. Howard (1998); Lindqvist u. Jonsson (1997)). Attacks are complex and may consist of several combined attacks. Therefore it should be allowed to categorize attacks into multiple categories to describe them in every possible detail, as also described in Bishop (1999).
- *Internal vs. External Threats* (Amoroso, 1994): The taxonomy concentrates on internal threats since external threats are widely covered by existing taxonomies regarding computer system attacks.

3.2.2 Applying Taxonomy to Related Research

An excerpt of the Socialbot attacks described in chapter 2 served as empirical examples to create the taxonomy to ensure quality, as for example also done in Alvarez u. Petrovic (2003). The attacks can only be categorized along dimensions if information is available in the literature. Therefore, a majority of the attacks only cover some of the taxonomic dimensions since no exhaustive description of the socialbot attacks is available. Some categories of the taxonomy are of rather theoretical character wherefore no example could have been found so far. However, they are part of the taxonomy to provide an exhaustive categorization system. A brief explanation of available information regarding the categorized socialbot attacks is followed by the overview of the empirical examples shown in Table 3.1. Three socialbot attacks for which extensive information is available (web, 2011; Boshmaf u. a., 2011; Thomas u. a., 2012) are categorized and explained in detail afterwards, they are also included in the overview given in Table 3.1.

- *Suspended Accounts on Twitter* (Thomas u. a., 2011): An overview of the attack is given in section 2.1.1. Large-scale spam attacks are reviewed by analyzing more than 1.1 million suspended accounts. The work does not reveal any information regarding the targets, therefore a categorization whether individual entities or entity collections were chosen is not available. A majority of investigated accounts were fraudulent accounts created by spammers exploiting the vulnerability that Twitter allows unverified account creation.
  The attacking accounts were hijacking trending topics and made use of unsolicited mentions by using API clients. One attack strategy was to flood the OSN with tweets before suspension, but no information is available if it was a distributed attack using several IP addresses or not. However, it shows that unrestricted user account actions were possible at least for a couple of days since those accounts were suspended after three days on average.
  No information is available about how other Twitter users reacted on the attack and if they took part in the affiliation attacks. A majority of the attacking accounts had less than ten followers. However, the social graph was slightly changed. No information whether the accounts were actively trying to retrieve followers or not is available.

Table 3.1: Taxonomy categories referring to empirical examples.

Dimension	Category			Empirical Examples
Target	Individual Entity			Aiello u. a. (2012); Boshmaf u. a. (2011)
	Entity Collection	by Social Relations		Nanis u. a. (2011); Boshmaf u. a. (2011)
		by Traits		no example found
		by States		Thomas u. a. (2012); web (2011)
	OSN			no example found
	Outside OSN			out of scope
Account Type	No Account			no example found
	Creeper			*pay4tweet*[10],*Pay with a Tweet*[11] (page 38)
	Compromised			Gao u. a. (2010); Grier u. a. (2010)
	Fraudulent			Thomas u. a. (2011); Aiello u. a. (2012); Lee u. a. (2011); Nanis u. a. (2011); Thomas u. a. (2012); Boshmaf u. a. (2011); web (2011), *Buy bulk accounts*[12] (page 38)
Vulnerability	CVE and others			out of scope
	OSN Functionality	System	Entity Information	Aiello u. a. (2012); Nanis u. a. (2011); Boshmaf u. a. (2011); web (2011)
			API Functionality	Thomas u. a. (2011); Lee u. a. (2011); Nanis u. a. (2011); Thomas u. a. (2012); Boshmaf u. a. (2011); web (2011); Grier u. a. (2010)
			IP Usage	Thomas u. a. (2012)
			Unverified Account Creation	Thomas u. a. (2011); Aiello u. a. (2012); Lee u. a. (2011); Nanis u. a. (2011); Thomas u. a. (2012); Boshmaf u. a. (2011); web (2011)
			User Account Actions	Thomas u. a. (2011); Boshmaf u. a. (2011)
			Channel or Topic Usage	Thomas u. a. (2011); Lee u. a. (2011); Gao u. a. (2010); Thomas u. a. (2012); Jindal u. Liu (2008); Grier u. a. (2010)
			User Interaction	Thomas u. a. (2011); Nanis u. a. (2011); Gao u. a. (2010); Boshmaf u. a. (2011); web (2011); Grier u. a. (2010)
		User	Participation in Malicious Actions	*pay4tweet*[10],*Pay with a Tweet*[11] (page 38)
			Social Relations	Aiello u. a. (2012); Lee u. a. (2011); Nanis u. a. (2011); Boshmaf u. a. (2011); web (2011)
			Communication	Aiello u. a. (2012); Nanis u. a. (2011); web (2011)
			Accepting Unknown Sources or Links	Grier u. a. (2010)
Attack Method	Within OSN Functionality	Abusive Usage of Topics		Thomas u. a. (2011); Grier u. a. (2010); Lee u. a. (2011); Thomas u. a. (2012); Jindal u. Liu (2008)
		Unsolicited Communication		Thomas u. a. (2011); Grier u. a. (2010); Nanis u. a. (2011); Gao u. a. (2010); web (2011)
		Attacking Social Graph		Nanis u. a. (2011); Boshmaf u. a. (2011); web (2011)
		Spoofing		no example found
		Leaving Traces		Aiello u. a. (2012)
		Setting Traps		Lee u. a. (2011)
	Outside OSN Functionality			out of scope
Results	Actively Changed OSN	Changed Social Graph		Thomas u. a. (2011); Aiello u. a. (2012); Lee u. a. (2011); Nanis u. a. (2011); Boshmaf u. a. (2011); web (2011)
		Modified Communication Channels or Topics		Thomas u. a. (2011); Aiello u. a. (2012); Lee u. a. (2011); Nanis u. a. (2011); Gao u. a. (2010); Thomas u. a. (2012); web (2011); Jindal u. Liu (2008); Grier u. a. (2010)
		Injured OSN		no example found
		Injured Account		no example found
	Passive Results	Gathered Sensible Information		Boshmaf u. a. (2011)

- *Spam Accounts on Twitter* (Grier u. a., 2010): Accounts used to send spam were analyzed. A majority of investigated accounts are reported to be compromised accounts tricking users to click on spam links. A much higher success rate for tricking users to click on spam links than known from email spam is reported. Information about which target groups are attacked is not given.
  More than three million tweets including spam URLs were identified. Based on missing restrictions regarding who can retweet tweets and interact with users, attackers were able to retweet users and add spam URLs to the tweets. The second strategy of the attackers was to create spam tweets combined with a trending topic hashtag. The attackers exploited exhaustive API functionality for performing Twitter actions.
- *Attacking book lovers on aNobii* (Aiello u. a., 2012): This socialbot attack was performed on aNobii an OSN for book lovers. Since aNobii was a rather small OSN when the attack was performed all users were included. Therefore, all individual entities were attacked. Two socialbots were created especially for the attack without any account verification. The bots exploited that entity information was unrestricted as well as they aroused targets' interest by leaving their traces on the targets' profiles. The targets started communicating with and following the socialbot.
- *Setting Traps on Twitter* (Lee u. a., 2011): The research study shows how honeypots were used to attract malicious users. This experiment is categorized as attack since the socialbots mimicked humans and pursued obscure goals. However, it has to be mentioned that no adversarial objective was pursued. The honeypots were created in a way to only attract malicious users, no information whether targets were organized as individual entities or entity groups is available. The honeypots used different strategies for creating tweets, one of them was to use the trending topics.
- *Socialbot Challenge on Twitter* (Nanis u. a., 2011): Researchers organized an experiment on Twitter, where socialbots were launched to study how well socialbots can embed in a chosen target group and to what extent they can be used for shaping the social graph of an OSN, described in detail in section 4.1.2.
  Nine socialbots were created to target users chosen by social relations. This was only possible since entity information was unrestricted and the social graph could be crawled. The socialbots were using exhaustive API functionality for following and communicating with the targets.
  Based on unrestricted user interaction the bots mentioned and retweeted targeted users in their tweets, they even created recommendation tweets recommending two or more users to each other. Users were vulnerable to the attack and communicated with or followed the socialbots.
- *Investigating Spam Accounts on Facebook* (Gao u. a., 2010): In this study wall messages on Facebook were investigated. More than 97% of investigated accounts were compromised accounts. The attacking accounts wrote messages to the targets'

Facebook walls.

- *Opinion Spam on amazon.com* (Jindal u. Liu, 2008): Reviews and reviewers from amazon.com have been analyzed to study opinion spam. The research does not reveal enough information about the account types involved in the attacks, nor about the targets to categorize the attack towards these dimensions. Attackers exploited the unrestricted topic or channel usage. The attack was conducted by abusive topic usage in the form of writing spam reviews for a specific product or topic.
- *Underground Markets*: Several websites offer services where Twitter users can sell single tweets. Some related research described the usage of underground-markets (Thomas u. a., 2011, 2012; Yang u. a., 2011a). Since no explicit information is available which underground-markets were used in the attacks, two samples for tweet selling *pay4tweet*[10] or *Pay with a Tweet*[11]as well as a sample for buying Twitter accounts in a bulk *Buy bulk accounts*[12] are used as examples for categorization.

Some socialbot attacks, for which detailed information is available are categorized by the taxonomy as follows.

Socialbot Challenge on Twitter

First, the taxonomy is applied to the socialbot experiment from web (2011) as shown in Table 3.2. Although socialbots did not show adversarial behavior this Twitter challenge can be categorized as socialbot attack since the socialbots mimic human behavior and humans as well as pursue obscure latent goals. An overview of the attack can be found in section 2.1.1.

A detailed explanation of the challenge and the resulting dataset can be found later on in section 4.1.1. This allows categorizing the attack along all dimensions. The socialbot attack was based on two strategies. First, socialbots followed the targets hoping for the targets to follow them back by using the auto follow back feature or by arousing interest. Second, socialbots extensively tweeted and also retweeted, replied to and mentioned the targets. The bots did not author a great deal of tweets in a short time frame to avoid suspension by the OSN providers.

Socialbot Experiment on Facebook

In the following the taxonomy is used to categorize a socialbot experiment conducted on Facebook in 2011. The study is explained in great detail in Boshmaf u. a. (2011). This

[10] `http://www.pay4tweet.com/`
[11] `http://www.paywithatweet.com`
[12] `https://buyaccs.com/en/`

Table 3.2: Applying the taxonomy to a socialbot attack conducted on Twitter in the beginning of 2011. Detailed information about the attack as basis for categorizing it is retrieved from web (2011).

Dimension	*Category*	*Description*
Target	Entity Collection by States	Targets were chosen based on the common characteristic that they were cat lovers.
Account Type	Fraudulent Accounts	Three main socialbots were created, partially supported by support bots.
Vulnerability	Entity Information	Twitter does not protect user information such as bio, the social graph or tweets, this was used to collect data about targets' interest.
	API Usage	The Twitter *REST* API[13] requires authentication for all requests (in the current version) and limits the number of API requests per account. Also the Streaming API[14] has rate limits. Nevertheless a broad API functionality is given by Twitter which was exploited by the socialbots.
	Unverified Account Creation	No verification is necessary for account creation on Twitter.
	User Interaction	Interaction between users is not restricted by Twitter by default. The socialbots were able to follow, retweet, mention and reply to targets in an unrestricted manner.
	Social Relations	Several users were susceptible to follow back the socialbots and therefore help the socialbots to successfully nest into the social graph.
	Communication	Several targets communicated with the socialbots.
Attack Method	Unsolicited Communication	Bots retweeted, replied to and mentioned targets in their tweets.
	Attacking the Social Graph	Socialbots followed targets, therefore tried to use auto follow back or arouse interest.
Results	Changed Social Graph	Bots established follow links to and from the targets.
	Modified Communication Channels or Topics	Bots retweeted, replied to and mentioned targets in their tweets and vice versa.

experiment is rated as attack although the socialbots did not show adversarial behavior since the socialbots mimicked human behavior and humans as well as pursued obscure latent goals.The attack is briefly reviewed in section 2.1.1. It was mainly based on the strategy to befriend strangers in a first step and befriend friends of successfully targeted users in a next step.

The taxonomic description is shown in Table 3.4.The attack exploited lots of vulnerabilities and pointed out that the Facebook Immune System (described in section 2.1.2) was not prepared for this kind of attack. For instance since the Facebook API offers a broad functionality and a large part of the user account information is unrestricted by default the attackers were able to crawl a lot of information about the targets and the social graph. Trust is partially transitive in Facebook, which means that friends of friends of a user are able to retrieve more information about this user than strangers. This may seem plausible under some circumstances. However, it also allows collecting data of users if their friends befriend strangers carelessly.

Although Facebook offers its users to restrict their privacy settings, the system generally enables public profiles and user data collection. Moreover privacy settings and Facebook functionality have changed repeatedly in the past which may have complicated securing a users' information. Table 3.3 gives an overview of data that could have been crawled by the attackers before and after the attack. One can see that on average 21.6% more information could have been crawled after the attack from users that directly befriended the socialbots. The extended neighborhood, which are basically friends of friends, revealed about 8.3% more information to the attackers than before the attack.

Table 3.3: Data that could have been crawled before and after the socialbot experiment conducted on Facebook copied from (Boshmaf u. a., 2011).

Neighborhoods	*Direct (%)*		*Extended (%)*	
Profile Info	*Before*	*After*	*Before*	*After*
Gender	69.1	69.2	84.2	84.2
Birth date	3.5	72.4	4.5	53.8
Married To	2.9	6.4	3.9	4.9
Worked At	2.8	4.0	2.8	3.2
School Name	10.8	19.7	12.0	20.4
Current City	25.4	42.9	27.8	41.6
Home City	26.5	46.2	29.2	45.2
Mail Address	0.9	19.0	0.7	1.3
Email Address	2.4	71.8	2.6	4.1
Phone Number	0.9	21.1	1.0	1.5
IM Account ID	0.6	10.9	0.5	23.7
Average	**13.3**	**34.9**	**15.4**	**23.7**

Table 3.4: Applying the taxonomy to socialbot attack conducted on Facebook in 2011. Detailed information about the attack as basis for categorizing it is retrieved from Boshmaf u. a. (2011).

Dimension	*Category*	*Description*
Target	Individual Entity	Randomly chosen user accounts were targeted in the first place.
	Entity Collection by Social Relations	In a second step users socially related to susceptible users were targeted.
Account Type	Fraudulent Accounts	102 accounts were created for the attack.
Vulnerability	Entity Information	Since user information regarding the social graph was available also friends of susceptible users could have been attacked.
	API Functionality	The Facebook's Graph API was used for performing social interactions.
	Unverified Account Creation	Facebook does not require verification for account creation.
	User Account Actions	Facebook usually requires a user to be logged in for API usage. However, this access control is insufficient since attackers can create applications that fetch permanent *OAuth2.0* (open standard for authorization[15]) tokens allowing API usage without login. Despite Facebook uses CAPTCHAs if an account shows high activity, the socialbots were not detected, since they adopted their behavior to avoid CAPTCHA solving Boshmaf u. a. (2011).
	User Interaction	Since Facebook offers the possibility that friend requests can be sent in an unrestricted way, attackers sent friendship requests to strangers.
	Social Relations	On average 35, 7% of the targeted users were susceptible to accept friend requests from the socialbots. Analyzing the data, researchers could show that even more relationships could be established (up to 80%) by increasing number of mutual friends between targeted users and the socialbots.
Attack Method	Attacking Social Graph	Information about socialbots sending friendship requests is available.
Results	Changed Social Graph	3,055 connections between socialbots and targets were established.
	Gathered Sensible Information	1,085,785 profiles could be crawled based on established social relations. See Table 3.3 for a comparison which amount of data could have been crawled after the attack.

Socialbot Attack strongly related to the Russian Election

A socialbot attack strongly related to the Russian election in 2011 (Thomas u. a., 2012) is categorized below. The attack is described in detail in section 2.1.1. This is a socialbot attack since it clearly shows adversarial behavior and socialbots pursued obscure goals. Attackers flooded hashtags which were used to organize protests regarding the Russian election by combining them with meaningless text until the hashtags were useless. It seems that attackers bought a bunch of fraudulent accounts at an underground-marketplace. This assumption is made based on the fact that the research found a large amount of other malicious or so far *silent* accounts showing same naming patterns. No information is available if the tweets were authored via the web interface or by exploiting API functionality. Table 3.5 shows how the attack can be categorized using the proposed taxonomy.

Table 3.5: Applying the taxonomy to a socialbot attack strongly related to Russian election conducted on Twitter in 2011. Detailed information about the attack as basis for categorizing it is retrieved from Thomas u. a. (2012).

Dimension	*Category*	*Description*
Target	Entity Collection by States	Top hashtags strongly related to the Russian election were targeted.
Account Type	Fraudulent Accounts	Approx. 25,860 fraudulent accounts probably out of a pool of 975,283 accounts offered by underground marketplaces were used.
Vulnerability	IP Usage	Although Twitter uses IP restriction algorithms and more than 110,189 different IP addresses were used by identified adversarial Twitter accounts, 39% of the used IP addresses were found to be blacklisted. This means that although Twitter generally uses IP restriction algorithms the system was vulnerable to the usage of blacklisted IP addresses.
	Unverified Account Creation	No verification is necessary for account creation on Twitter.
	Channel or Topic Usage	Unrestricted hashtag usage was exploited.
Attack Method	Abusive Usage of Topics	Tweets were created using hashtags in combination with unrelated content.
Results	Modified Communication Channels or Topics	Meaning of specific hashtags was changed or destroyed by 440,793 tweets.

4 Empirical Studies

This part of the thesis addresses *Research Question 2* and *Research Question 3* (see section 1.1). It shows two different studies regarding the impact of socialbot attacks and behavior of targeted users. Two datasets are available for the experiments. The first experiment regarding the susceptibility of users is performed on both datasets. The second study regarding the potential impact of socialbot attacks is performed on the second dataset. Below, datasets are explained.

4.1 Socialbot Challenges and resulting Datasets

Datasets from two different socialbot challenges performed on Twitter were used for the following empirical studies. This work concentrates on investigating the success of socialbot attacks and potential impact of socialbots based on the two existing datasets. The author of this thesis and colleagues were not involved in nor did participate in the design, setup or execution of these challenges. This chapter first describes the socialbot challenges and their resulting datasets followed by the description and results of the experiments conducted on the datasets.

4.1.1 Dataset 1 from the WebEcology Challenge

This Socialbot Challenge was performed with the objective to explore reactions of targeted users when socialbots interact with them. It was organized by Tim Hwang and the Web Ecology Project (WebEcology) and took place in January and February 2011 (web, 2011). Competing teams were developing socialbots with the objective to successfully interact with targets, i. e. make targets reply to, mention or retweet the bots or create links. The target group consisted of 500 unsuspecting Twitter users which were selected by a common characteristic – all users had an interest in or tweeted about *cats*. The majority of targets exhibited a high activity level, that means they tweeted more than once a day.

Every team was represented by one lead socialbot (the only socialbot allowed to score points) and an arbitrary number of support bots. Participating teams collected points

for every successful interaction between their lead socialbots and users within the target group. One point was awarded for targets following a lead bot and three points were awarded for targets replying to, mentioning or retweeting a lead bot.

The teams had to follow a set of rules[1]. First, the socialbots had to act as autonomous agents since human interaction was forbidden during the challenge. Participating teams were not allowed to report each other as spam to the OSN provider. However, they were allowed to use other strategies and countermeasures to harm each other. Teams and their bots were not allowed to unveil the existence of the challenge to anyone outside the group of challenge members during the competition. Teams had to be willing to provide their source code for the bots under an open source license after the challenge took place. Collaboration between teams was allowed and supported by challenge organizers.

After a developing period of 14 days, the game started on the January 23rd 2011 (day 1) and ended February 5th 2011 (day 14). During this period, socialbots were autonomously active for the first 7 days. At the 30th of January (day 8) the teams were allowed to update their source code and change strategies. After this optional update the socialbots continued to act autonomously for the remaining time of the challenge.

The following three teams competed in the challenge.

- *Team A – @sarahbalham* The bot mimicked a young woman that grew up on the countryside, had just moved to the city and was looking for friends. This team did not construct a socialbot-network, they only used the lead bot. @sarahbalham authored 143 tweets which is rather low in comparison to the other teams and used only a few @replies and hashtags. Despite low activity level this team could collect the highest number of followers - 119 users followed sarahbalham. Overall the team was only able to collect 170 points since only 17 interactions with targets were counted.
- *Team B – @ninjzz* The woman impersonated by this socialbot did not provide much personal information only that she was a bit shy and looking for friends on Twitter. Ninjzz was supported by 10 other socialbots which also created some tweets. This socialbot was rather defensive in the beginning but changed the strategy on day 8 and acted in a much more aggressive way in the second part of the challenge. Overall this team created 99 mutual connections and 28 interactions what resulted in 183 points.
- *Team C – @JamesMTitus* The socialbotclaimed to be a 24 old guy from New Zealand, new on Twitter and a real cat enthusiast. Team C and their socialbot JamesMTitus won the game by collecting 701 points with 107 mutual connections and 198 interactions. This team had five support bots that only created social connections but did not tweet at all. The team picked a very aggressive strategy

[1] `http://robotandhwang.com/Socialbots/Public%20Socialbots%20Rules.pdf`

where the bot tweeted a lot and also made extensive use of @replies, retweets and hashtags nevertheless, they managed to avoid suspension.

The dataset was provided by the Web Ecology Project as a *MySQL* data dump. Table 4.1 provides a basic description of the dataset.

Table 4.1: Description of the Social Bot Challenge dataset

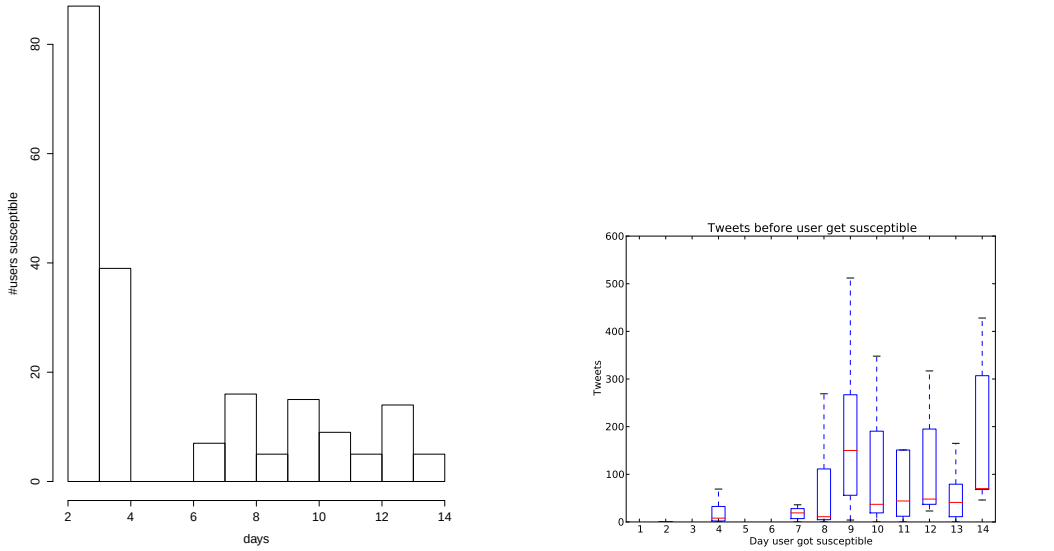
Leadbots	3
Susceptible Users	202
Non-Susceptible Users	298
Mean Nr of Tweets per User	146.49
Mean Nr of Follower/Followees per User	8.5

Figure 4.1a shows infections over time, i. e. it depicts on which day of the challenge targets interacted with socialbots for the first time. One can see from this figure that at the second day of the challenge already 87 users had become susceptible. One possible explanation for this might be the usage of auto-following features which some of the targets might have used. One can see from Figure 4.1b that for the users who became susceptible at an early stage of the challenge just a few tweets are available in the dataset. This is a limitation of the dataset which includes only tweets authored between the 23th of January and the 5th of February and social relations which where existent at this point in time or created during this time period. Due to crawling limitations of Twitter that only the 3,200 most recent tweets per user can be collected via the API it was not possible to retrieve tweets from before the challenge for this work since timespan between receiving this dataset and the competition was too large.

4.1.2 Dataset 2 from the PacSocial Challenge

This dataset was provided by the Pacific Social Architecting Corporation (PacSocial) a corporation focusing on technologies such as socialbots to investigate possible influence on shaping the structure of online social networks and communities in a large-scale. PacSocial performed an experiment on Twitter at the end of 2011 from which the dataset was collected.

The experiment was designed to consist of a *control phase* (ctr) and an *experimental phase* (exp). The control phase lasted 33 days where ongoing developments were not influenced in any way or at any time no socialbots were launched during this period. It was solely used to capture information and data about tweets and development of the social graph over time. Information was collected for 2,700 users which were structured in two observation groups: one consisting of 1,800 and another one consisting of 900 users which were partially connected through follow links. The exact way in which those users were chosen is unknown. The control phase was immediately followed by a 21 day



(a)Number of users interacting with socialbots for the first time

(b)Time of infection and number of tweets until infection

Figure 4.1: Number of infected users and time of infection.

experimental phase. The two user groups from the control phase were restructured into nine equally sized target groups consisting of 300 users before the experimental phase started. In this phase, nine socialbots were released – one bot per target group. Hence, there is additional follow information available for the control phase. The restructure did not show any other consequences since bots were launched afterwards.

The socialbots were all designed in the same way but the level of the daily interaction and activity of the socialbots was chosen in a random manner from a possible pool of choices and within given boundaries. The major purpose of the socialbots was to act as recommender bots. They should try to actively influence social link creation between human users by recommending them to each other. Additionally the socialbots should simply interact with users within their target groups to gain followers and establish as many conversations with the targets as possible.

The dataset was provided by the PacSocial Group as a *MySQL* data dump. It contains tweets which were published after the control phase by any of the socialbots or target users when communicating with a socialbot, i. e. if they were replying to, mentioning or retweeting socialbots. Additionally, the dataset contains follow information within the two observational groups during control and within the nine target groups for the experimental phase. This means that some essential information was missing in the dataset:

1. The targets' tweeting behavior during the control phase was missing.
2. The targets' tweets created during the experimental phase which did not show any interaction with the socialbots (i.e. reply to, mention or retweet) were missing.

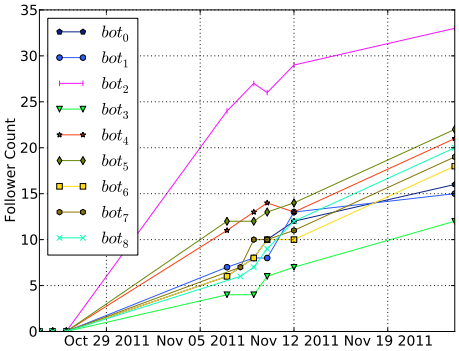
To receive this additional and essential information a *Crawler* was written using *Python*[2] to collect tweets. The crawler used the Twitter API to receive JSON formatted tweet information. Because of Twitter restrictions the maximum number of available tweets per user via the Twitter API is 3,200. Therefore, all available tweets starting about two months prior to the control period until the end of the control period but only up to 3,200 tweets per user were collected additionally. One has to notice that since the data was not crawled immediately after the challenge, it was not possible to collect all tweets for all users for the period starting two months before the challenge. Table 4.2 gives an overview of the dataset.

Table 4.2: Dataset description

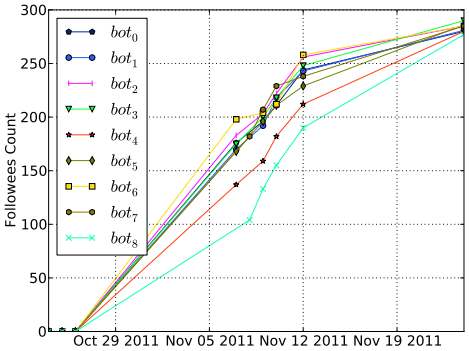
Socialbots	9
Targets	2,700
Targets following Bots	192
Targets communicating with Bots	232
Nr of Tweets	1,004,048

The number of incoming and outgoing follow links of the socialbots within their target groups is shown in Figure 4.2. Values are plotted over time based on available snapshots where values measured at specific moments in time are indicated by markers. As opposed to the control period where snapshots of the social graph are available on an almost daily basis the experimental phase only consists of snapshots from the first and the last few days. Information of about 11 days during the experiment (from 26th of October until 6th of November) is missing. This may be a limitation of the dataset. Since the socialbots all act based on the same strategy using random values within given boundaries a similar evolution of the social graph of the socialbots would be expected. Figure 4.2a shows that a majority of the socialbots attracts approximately the same number of users but bot_2 can collect about twice as much followers during the first days than the others. Comparing to the number of followees per socialbot as shown in Figure 4.2b this cannot be explained by an extraordinary follow behavior of the socialbot. Also the number of tweets authored per socialbot does not show any abnormality for this socialbot as shown in Figure 4.3a. Tweets authored by targeted users in which they replied to, mentioned or retweeted socialbots are shown in Figure 4.3b per socialbot and also show higher values for the highly followed bot_2. Since all bots pursued the same strategies and no abnormality in the behavior of bot_2 was measurable no reasons for those outlying results could be found.

[2] a modern, interpreted, high-level programming language `http://www.python.org/`

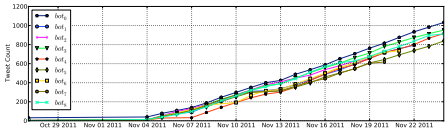


(a)Number of Followers per socialbot over time

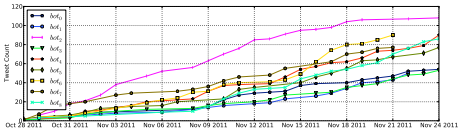


(b)Number of Followees per socialbot over time

Figure 4.2: Incoming and outgoing links per socialbot



(a)Tweets authored by socialbots



(b)Tweets replying to, mentioning or retweeting socialbots authored by targets

Figure 4.3: Tweets created by or referring to socialbots

4.2 Research Question 2: Susceptibility of Users

In the following, an experiment according to the second research question – *To what extent is it predictable whether a user will be susceptible to a socialbot attack or not? Do susceptible users show any specific characteristics which allow to differentiate them from non-susceptible users?* – is described. To perform this study both datasets described in 4.1 are used. The experiment on the WebEcology dataset was conducted in collaboration with two colleagues and is already published in Wagner u. a. (2012). It is investigated if and to what extent a susceptible user can be predicted based on different features calculated for each user. A user is defined to be susceptible if he or she interacts with one of the socialbots in one of the following ways: follows a socialbot, mentions, retweets or replies to a socialbot. Different machine learning estimators are used to train models for binary classification to distinct between susceptible and non-susceptible users. The two studies are performed in a similar but not identical way since the two competitions did have different objectives and led to differently organized datasets.

4.2.1 Feature Engineering

To this end, three feature sets are identified that can be leveraged to identify susceptible users: *linguistic, behavioral* and *network features*. Features are calculated based on user streams which are the sum of available tweets per user and a follow network snapshot.

For the WebEcology dataset a user stream is created by combining all tweets a user authored up to the time he or she became susceptible or until the end of the study if the user did not become susceptible. For the follower network the first available snapshot of social relations recorded at the 26th of January 2011 (day 4) is used.

For the PacSocial challenge only information from the control phase is used to ensure that information used to predict susceptible users is not influenced or biased by socialbots or any future information. User streams are formed by tweets from about 2 months prior the control period (max. 3,200 tweets) until the end of the control period (including 21st of October 2011). The follow network snapshot is crawled on the 22nd of October 2011, directly before socialbots were launched. Below, the features are described in detail.

Linguistic Features

In order to analyze the sentiment of a user's tweets a computer based text analysis program is used called *Linguistic Inquiry and Word Count* (LIWC), which is described in Tausczik u. Pennebaker (2010). LIWC analyzes a given text by counting *search words*. Defined words and word stems are used as search words (Pennebaker u. a., 2003). They have been categorized into over 21 linguistic superdimensions which can be expanded to 70 linguistic subdimensions. These dimensions include *standard language categories* (e. g. articles, prepositions, pronouns including first person singular, first person plural, etc.), *psychological processes* (e. g. positive and negative emotion categories, cognitive processes such as use of causation words, self-discrepancies), *personal concerns* (e. g. work, home, money, achievement) and *spoken categories* (assent, nonfluencies and fillers).

In this work the linguistic dimensions[3] are used as linguistic features depending on what performed best – super- or subdimensions were used. They are computed based on the aggregation of tweets authored by each target. The calculation of the LIWC features was run by a colleague. Linguistic features are not described in detail but the ones that seem to be relevant for modeling the susceptibility of users are elaborated in greater detail in the result sections 4.2.3 and 4.2.4 and the full list of LIWC features can be found in the Appendix.

[3] `http://www.liwc.net/descriptiontable1.php`

Behavioral Features

These measures are calculated based on tweeting behavior of the targets. Features are largely based on features introduced in Wagner u. Strohmaier (2010). Behavioral measures can be used to characterize user streams and reveal structural differences between them. Below, measures are explained and usage for gauging the susceptibility of users is elaborated.

Message Count
Number of tweets a user created during the observation period.

Question Ratio
Number of questions asked by a user in his tweets.

Conversational Features
Conversational messages are tweets in which a user replies to, mentions or retweets another user. Following variables have to be defined for calculating conversational features. A user stream M is the set of messages a user created. For each user stream the set of unique users in this stream U_M is defined with values from $1..n$ where n is $|U_M|$, the number of different users. M_c is the set of conversational messages in a user stream.

- *Conversational Variety*: CV represents the average number of different users per tweet message in a stream, that is:

$$CV = \frac{|U_M|}{|M|}. \tag{4.1}$$

  A high conversational variety indicates that a user talks to many different users.

- *Conversational Balance*: To quantify the conversational balance of a stream an entropy-based measure is defined which indicates how evenly balanced the communication efforts of a user are distributed across his communication partners. If a user has no communication partner the conversational balance is defined to be zero. Otherwise is is calculated by using the normalized entropy:

$$CB = -\frac{\sum_{u \in U_M} P(u|M) \cdot log_2(P(u|M))}{log_2(|U_M|)}. \tag{4.2}$$

  The entropy measure is normalized by the logarithm of the number of users in the stream. This way, entropy measures can be compared independently of how conversational a user's tweets are. Therefore, a high conversational balance indicates that the user talks equally much with a given set of users. A high score indicates that it is hard to predict with whom a user will talk next. Conversely a low conversational balance indicates that a user talks with few selected users in most of his tweets.

- *Conversational Coverage*: This measure indicates the proportion of messages in a user stream that are conversational:

$$CC = \frac{|M_c|}{|M|}. \tag{4.3}$$

Topical Features

Several surrogate measures can be used for topics, such as the result of automatic topic detection or manual labeling methods. For this work, hashtags are defined as topic surrogate therefore tweet messages containing any hashtags are defined as topical messages. Variety, balance and coverage are calculated following the formula for conversational features but using hashtags instead of users where the set of unique hashtags is defined by R_h and topical messages by M_h.

- *Topical Variety*: It indicates how many different topics the user is interested in. It is the average number of hashtags per message in a user stream:

$$TV = \frac{|R_h|}{|M|}. \tag{4.4}$$

- *Topical Balance*: The balance is defined as the normalized entropy of the probability distribution of hashtags in the messages of the user stream. If a user stream does not contain any hashtags, the balance is zero. Since values are normalized a high topical variety indicates an equal distribution of interest.

$$TB = -\frac{\sum_{h \in R_h} P(h|M) \cdot log_2(P(h|M))}{log_2(|R_h|)}. \tag{4.5}$$

  A low entropy indicates a low uncertainty in predicting the next topic used in the user's tweet.

- *Topical Coverage*: The topical coverage indicates how many of the created tweets contain hashtags.

$$TC = \frac{|M_h|}{|M|}. \tag{4.6}$$

Informational Features

A message is called an informational message if it contains a URL. The feature calculation is similar to conversational and topical features but using URLs instead of users or hashtags. The set of URLs is defined as R_i, the set of informational messages is denoted by M_i.

- *Informational Variety*: It indicates how many different URLs a user tweets normalized by the total number of tweets:

$$IV = \frac{|R_i|}{|M|}. \tag{4.7}$$

- *Informational Balance*: If the user stream does not contain any URLs the balance is zero.
$$IB = -\frac{\sum_{i \in R_i} P(i|M) \cdot log_2(P(i|M))}{log_2(|R_i|)}. \tag{4.8}$$
Again a high balance indicates a high uncertainty in predicting the next URL whereas a low balance indicates higher predictability.

- *Informational Coverage*: Shows proportion of all tweets to the tweets including an URL.
$$IC = \frac{|M_i|}{|M|}. \tag{4.9}$$

Lexical Features
Features regarding keywords in user streams are measured. The user stream M is modified by eliminating stopwords so that it only consists of a set of keywords R_k. Again, M denotes the set of all messages in the user stream.

- *Lexical Variety*:
$$LV = \frac{|R_k|}{|M|}. \tag{4.10}$$
A high lexical variety indicates that a user talks about different topics and/or has a large vocabulary.

- *Lexical Balance*: The lexical balance measures the uncertainty of the next keyword of the user:
$$LB = -\frac{\sum_{k \in R_k} P(k|M) \cdot log_2(P(k|M))}{log_2(|R_k|)}. \tag{4.11}$$

A high balance indicates a high uncertainty and conversely.

Temporal Features
Temporal features can be used to measure the temporal consistence of a user's tweeting behavior. For every tweet the hour of the creation time is extracted. R_t is the set of distinct creation times (hours).

- *Temporal Variety*: A high temporal variety indicates that a user tweets at different hours of a day:
$$TPV = \frac{|R_t|}{|M|}. \tag{4.12}$$

- *Temporal Balance*: The temporal distance shows the uncertainty of *when* the next tweet is created:
$$LB = -\frac{\sum_{t \in R_t} P(t|M) \cdot log_2(P(t|M))}{log_2(|R_t|)}. \tag{4.13}$$

A high uncertainty indicates that the user tweets at different hours of the day about equally often and it is hard to predict at which hour the next tweet will be created. A low value shows that a user almost always tweets at the same time.

Network Features

To include network structure in the predictions three directed networks are created by using all available users (also including users outside the network if available) as nodes:

- A *Follower Network* is created from the target follower structure on Twitter. A directed edge from target A to target B means that target A is following target B.
- A *Retweet Network* is a network based on the targets' retweets. If user A retweets user B a directed edge from user A to user B exists. For this network also retweeted users outside the target groups are used as nodes (if available in the dataset) to calculate network measures in a next step.
- An *Interaction Network* is also based on the targets' tweets. If user A interacts with user B (i. e. replies to, mentions or retweets) a directed edge from user A to user B exists. This network also makes use of all users available from the tweets to build a network as comprehensive as possible.

Based on the three networks, the following network features are calculated:

In- and Outdegree
The indegree is the sum of incoming edges of a node whereas the outdegree counts outgoing edges per node. A high indegree indicates that a user has many followers (follower network), is often retweeted (retweet network) or other users like to interact with him by replying to him, retweeting or mentioning the user (interaction network). A high outdegree measures the users follow, retweet and interaction behavior.

Authority and Hub Score
Referring to the HITS algorithm, a precursor to the PageRank, introduced by Kleinberg (1999) (briefly described in the chapter 2) the *Authority* and *Hub* score is calculated for all targets. A high authority score indicates that a user has many incoming links from targets with a high hub score. Whereas a high hub score indicates that a user has a high outdegree to users with a high authority score. In other words, nodes with a high authority score tend to be important nodes whereas nodes with a high hub score tend to link to important nodes in a network.

Clustering Coefficient
The clustering coefficient describes how closely related the neighborhood of a node n is. It calculates the proportion of existing edges e between a node n and his neighbors k_n

in a graph G and is calculated per node in a directed graph as follows:

$$CC_i = \frac{|\{e_{j_k}\}|}{k_i(k_i-1)} : n_j, n_k \in N_G(n), e_{j_k} \in E(G) \tag{4.14}$$

4.2.2 Experimental Setup

First the datasets are cleansed by removing *bad users* which are users not available until the end of the competition for different reasons, e. g. because their accounts are deleted, suspended or restricted. Most of the calculated features require a certain amount of tweets in order to contain meaningful information about a user. Therefore, users which became susceptible before tweets authored by them could have been captured are removed. Next, the datasets are split into two classes: *susceptible users* (interacted with socialbots) and *non-susceptible users* (did not interact with socialbots). They were balanced, which means that an equal part of each class of users is used for training and test sets. Features are calculated for every user as described in section 4.2.1, except entropy based measures were not normalized for the study performed on the WebEcology dataset. The objective of the study is to identify features associated with susceptible users. For this purpose, several classifiers are trained to perform a binary classification task belonging to the class of *supervised learning* tasks where class labels are given for the data.

For the study on the WebEcology dataset (see 4.1.1), Python is used for preprocessing tasks, such as data cleansing and feature calculation. For classification the *R Project*[4] is used. For performing the experiment on the PacSocial dataset (see 4.1.2) Python is used for preprocessing and feature calculations, but also the classification task is performed using a Python library, *scikit-learn*[5]. The R package is only used for feature importance calculations with the *Boruta* package which is explained in more detail later on in section 4.2.4.

Model Selection

A few important terms regarding classification using machine learning algorithms are explained briefly to give a short introduction. Model selection is the process of determining the best hypothesis for a learning problem as described in Bishop (2006). Datasets are divided into *training, test* and *validation set*. First a model is learned on a training set to tune parameters of the estimator and validated on the validation set to decide which model performs best. The trained model is applied to a test set to measure how

[4] An open source software for statistical tasks which is commonly used in statistics and datamining communities for statistical calculations – `http://www.r-project.org/`

[5] open-source library for machine learning tasks – `http://scikit-learn.org/stable/`

well the model generalizes. Since usually training data only covers a small fraction of possible input it is an important goal that a model can generalize from training data for good performance on test data. Different *score functions* can be used to evaluate the performance of a model. Trainings, test and validation data should not overlap, e. g. if data used for testing is (partially) the same as training data results would be distorted.

Partitioning the data can be a challenging task since the right split proportion is crucial. The training set should be large enough to learn how to fit the data and avoid *overfitting* which means that the model cannot properly generalize from the training data. The validation set should be large enough to enable good model selection by predicting the true error and the test set must consist of enough data to enable a good final evaluation of the trained model.

If the dataset is rather small, a technique called *k-fold-cross validation* can be used to split data into training and test set. With k-fold-cross validation data are partitioned into k equally sized groups where $k-1$ groups are used for training and the remaining group is used for evaluating the trained models. This procedure is repeated for all k splits and scores are averaged over all runs to ensure better generalization and reduce variations. This way, all data can be used for training and evaluation but trainings and test sets do not overlap per run (Bishop, 2006). *Stratified-k-fold validation* can be used to maintain the same proportion of data for each class in the training and test sets.

To evaluate performance of a classifier different score functions can be used such as *Accuracy*, *Recall*, *Precision* and *F1-score*. They are based on the number of samples which are correctly classified in the positive class – true positives (TP), erroneously classified in the positive class – false positives (FP), correctly classified in the negative class – true negatives (TN) and erroneously classified in the negative class – false negatives (FN).

$$Accuracy = \frac{TP + TN}{|samples|} \tag{4.15}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{4.16}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{4.17}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{(Precision + Recall)} \tag{4.18}$$

To evaluate the values of the score function one has to calculate a *baseline* which indicates which classification result would be reached on average by using random features. The *Receiver Operating Characteristic* ROC curve is a method to visualize the prediction accuracy of a binary classifier by plotting the *true positive rate* (true positives out of all positives) against the *false positive rate* (false positives out of all negatives).

Classifier Description

Many different classifiers were used to train and test the model. Below, the best performing classifiers, which performances are shown later on in the result sections (see 4.2.3 and 4.2.4) are explained briefly:

- *K-Nearest-Neighbors* (knn): To classify an object the k closest neighbors from the training data to this object are identified. The object is than assigned to the class where most of the k neighbors belong to (Bishop, 2006).
- *Ensemble Methods*: Such methods leverage several models to reduce the error and improve robustness for a chosen algorithm. In general two types of ensemble methods are distinguished: *averaging* and *boosting* methods. Averaging methods use independent, strong classifier and average their results with the effect that the performance may be improved while the variance should be reduced. Boosting methods use sequential learning where different classifiers are trained in a sequence and use weight values of the previous classifiers. Therefore, classifiers with a weak performance can be combined to create a method with improved performance (see Bishop (2006) and scikit-learn[6]).
  - *Random Forests* (rfc): This is an averaging ensemble method developed by Breiman and Cutler. The forest uses several classification trees where each of the trees is drawn from a sample with replacement. A random subset of features is used to classify the object. The number of trees used for the forest plays an important role (see scikit-learn documentation[6] and Berkely documentation[7]). In the implementation used by the R project the forest then assigns the object to the class to which the majority of trees classified the object[7], whereas the scikit-learn implementation calculates probabilities that an object belongs to a class and chooses the class with the highest average probability[6].
  - *Gradient Boosting Models*: This ensemble method uses the boosting method. It builds the additive model by using several basis classifiers in a forward stepwise way. The defined loss function is minimized at each step. This way several weak learners can be combined to learn one strong model. For working with the R project the *Generalized Boosted Regression Model* (gbm) can be used (Ridgeway, 2005). The scikit-learn library offers the *Gradient Boosting Classifier* (gbc) (see scikit-learn documentation[6]).
- *Support Vector Machines* (svm): These classifiers do not use fixed basis functions, but rather use basis functions centered on the training data to ensure practical applicability for higher dimensional data, as described in Bishop (2006).

[6] `http://scikit-learn.org/0.12/modules/ensemble.html`
[7] `http://stat-www.berkeley.edu/users/breiman/RandomForests/cc_home.htm`

  - *Support Vector Classifier* (svc): A svc tries to find the largest possible distance between the training samples and the decision boundaries. Only vectors of training samples helping to describe these boundaries are used as *support vectors*. If no linear decision boundary can be found for the data, the so called *kernel trick* can be used. It basically allows non-linear classification by mapping the input vectors into a higher dimensional space where they are linearly separable (Bishop, 2006).
  - *Linear Support Vector Classifier* (linearSvc): This classifier is related to the svc, but only uses linear basis functions (see scikit-learn documentation[8]).
- *Elastic-net Regularized Generalized Linear Models* (glmnet): This is an algorithm to estimate generalized linear models by using different penalty methods, i. e. the *lasso*, *ridge regression* and the *elastic net*, a combination of the two before mentioned penalty methods (see detailed explanation in Friedman u. a. (2010)).
- *Partial Least Square Regression* (pls): This is a regression model trying to maximize covariance between the observed variables and the response variables (Mevik u. Wehrens, 2007).

4.2.3 Research Question 2 – Experiment on the WebEcology Dataset

This experiment was conducted in collaboration with two colleagues. As one can see in Figure 4.1b, no tweets are available for a majority of the users who became susceptible before day 7, therefore all users susceptible before day 7 are removed from the dataset to provide meaningful values. While this means 133 susceptible users cannot be used as samples for the experiments it seems that:

(i) the remaining 76 susceptible users and 298 non-susceptible users are sufficient to train and test classifiers
(ii) eliminating those users that might have used an auto-follow feature is a good decision since they are less interesting to study from a susceptibility viewpoint.

After data cleansing, 76 susceptible and 298 non-susceptible users are available for the prediction task. The dataset is divided in a balanced training and test set. To overcome the fact that the number of samples is relatively low and to average results a 10 fold-cross validation is performed. All features are used for the classification task performed F1-score, recall and precision are used as accuracy measures for evaluating classifiers' performance. The ROC curve is used as a ranking criterion for feature importance.

[8] `http://scikit-learn.org/stable/modules/svm.html`

Results and Evaluation

In the following, the results from the experiment regarding RQ 2 performed on the WebEcology dataset are discussed, followed by an interpretation of the results. The results are compared to a random baseline, which is given with an accuracy value of 50%, since balanced datasets are used.

Classifier Selection

Performance of the classifiers is shown in Table 4.3. Classifiers are explained briefly in section 4.2.2. One can see that generalized boosted regression models (gbm), elastic-net regularized generalized linear models (glmnet) as well as k-nearest neighbor (knn) perform best, with highest F1-score. In the following the fitness of the features for susceptibility prediction is investigated for the gbm as one of the best performing classifiers.

Table 4.3: Comparison of classifiers' performance for the WebEcology experiment. The random baseline is given by accuracy values of 0.50. Classifiers are described briefly in section 4.2.2.

	Susceptible			**Non-Susceptible**			Overall
Model	**F1**	**Rec**	**Prec**	**F1**	**Rec**	**Prec**	**F1**
random	0.50	0.50	0.50	0.50	0.50	0.50	0.50
gbm	0.71	0.70	0.74	0.70	0.74	0.68	0.71
glmnet	0.69	0.75	0.67	0.73	0.72	0.77	0.71
pls	0.67	0.69	0.68	0.68	0.71	0.70	0.68
knn	0.70	0.71	0.71	0.72	0.75	0.71	0.71
rf	0.68	0.72	0.66	0.70	0.70	0.74	0.69

Fitness of Features for Prediction

To understand which features are most predictive the importance of different features is explored by using the best performing model. Table 4.4 shows the importance ranking of features using the area under the ROC curve as a ranking criterion. One can see that the most important feature for differentiating susceptible and non-susceptible samples is the out-degree of a user's node in the interaction network.

Figure 4.4 shows the box plots for the top 20 features. Non-susceptible users are represented by yellow boxes and susceptible users by red boxes. One can observe differences of the feature values.

Interpretation of Results

Figure 4.4 suggests that susceptible users tend to actively interact with more users than non-susceptible users do on average. One can conclude that susceptible users tend to have a larger social network or communication network, respectively. One possible explanation for that is that susceptible users tend to be more active and open and therefore easily create new relations with users. Results also show that susceptible users

Table 4.4: Importance ranking of the top features (WebEcology experiment) using the area under the ROC curve (AUC) as ranking criterion. The importance value is proportional to the most important feature which has an importance value of 100%.

Feature	Importance
out-degree (interaction network)	100.00
verb	98.01
conversational variety	96.93
conversational coverage	96.65
present	94.66
affect	90.15
personal pronoun	89.71
first person singular	89.27
conversational balance	87.28
motion	87.28
past	86.56
adverb	86.20
pronoun	84.41
negate	84.33
positive emotions	83.25
third person singular	82.38
social	82.02
exclusive	81.86
auxiliary verb	81.70
in-degree (interaction network)	81.66

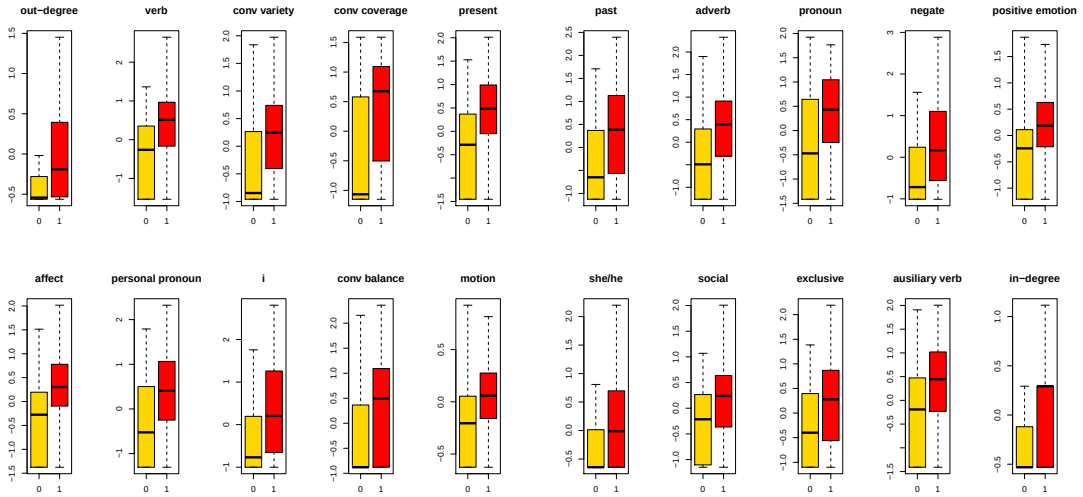


Figure 4.4: Box plots for the top 20 features (WebEcology experiment) according to the area under the ROC curve (AUC). Yellow boxes (class 0, left) represent non-susceptible users, red boxes (class 1, right) represent susceptible users. Differences between susceptible and non-susceptible users can be observed.

also tend to have a high in-degree in the interaction network which indicates that most of their interaction efforts are successful.

Further, it seems that susceptible users tend to use more verbs (especially present tense verbs but also past tense verbs and auxiliary verbs) and use more personal pronouns (especially first person singular but also third person singular) in their tweets. This suggest that susceptible users tend to use Twitter to report about what they are currently doing.

Interestingly, results also show that susceptible users have a higher conversational variety and coverage than non-susceptible users, which means that susceptible users tend to talk to many different users on Twitter with a high conversational purpose. This indicates that susceptible users tend to use Twitter mainly for a conversational purpose rather than an informational purpose. Further, susceptible users also have a higher conversational balance which indicates that they do not focus on just a few conversation partners but spend an equal amount of time in communicating with a large variety of users. It suggests again that susceptible users are more open to communicate with others also if they are not within their closed circle of friends.

Furthermore results suggest that susceptible users show more affection than non susceptible users indicated by an extensive usage of affection words - especially words exposing

- positive emotions, such as *love* or *nice*
- words indicating social affection, such as *mate* or *friend*
- motion words, such as *go*, *car*
- adverbs, such as *really*, *very*
- exclusive words, such as *but*, *without*
- negation words, such as *no*, *not*, *never*

It seems that susceptible users tend to use Twitter to talk about their activities and communicate on an emotional basis.

To summarize, findings suggest that susceptible users tend to use Twitter mainly for a conversational purpose (high conversational coverage) and tend to be more open and social since they communicate with many different users (high out-degree and in-degree in the interaction network and high conversational balance and variety), use more social words and show more affection (especially positive emotions) than non-susceptible users.

4.2.4 Research Question 2 – Experiment on the PacSocial Dataset

After removing susceptible users for which no tweets are available, 285 susceptible users remain in the dataset. A balanced dataset containing 285 susceptible and 285 non-susceptible samples is used for the classification. First, data is scaled, since some of

the classifiers require data that looks like standard normally distributed data otherwise they may perform badly. For example, Support Vector Machines used with nonlinear kernel (e. g. RBF kernels) expect features to be centered around zero and with variance in the same order. Otherwise, features with significantly higher variance than others could dominate the objective function and the model could not be trained correctly as described in the scikit-learn documentation[9]. The classification task is performed for different classifiers by applying meta optimization for parameters and stratified-k-fold-cross validation with $k = 4$ and $k = 10$.

Results and Evaluation

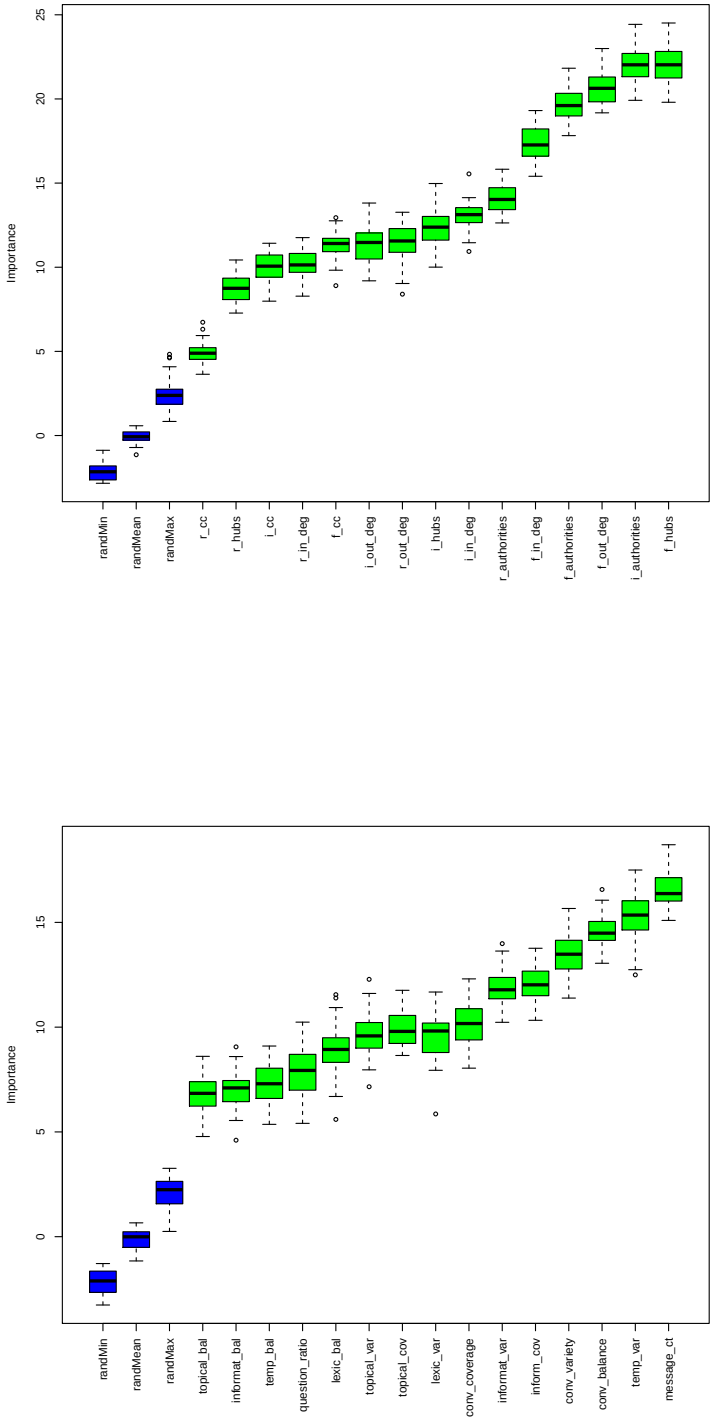
Next, the results from the experiment regarding RQ2 performed on the WebEcology dataset are discussed, followed by an interpretation of the results. The results are compared to a random baseline, which is given with an accuracy of 50% since balanced datasets are used.

Feature Selection

Since lots of features are calculated for the classification task and it is unknown in advance which features may be relevant and which not, feature selection is a common strategy. The *Boruta* package for *R Project* is used to perform statistical feature relevance tests, described in Kursa u. Rudnicki (2010). Tests are based on the *Random Forest* algorithm. For each feature an additional random feature is introduced by copying and then shuffling values from the corresponding feature. Then for each tree it is measured how the accuracy values decrease by using feature values randomly shuffled between objects. The feature importance measure is calculated by dividing the average decrease of accuracy by the standard deviation (Kursa u. Rudnicki, 2010)). Then, the classification is performed repeatedly to ensure statistical valid results by comparing all features to random features and calculating importance values.

One can see the results of the Boruta test for behavioral features in Figure 4.5a, network features in Figure 4.5b and linguistic features for all subcategories in Figure 4.6a as well as only supercategories in Figure 4.6b. Boruta results suggest that all calculated behavioral and network features can be used for classification since they perform better than random features. Comparing results for linguistic features suggests the usage of the 21 supercategories rather than the usage of all 70 subcategories since only about half of the subcategory features perform better than random features.

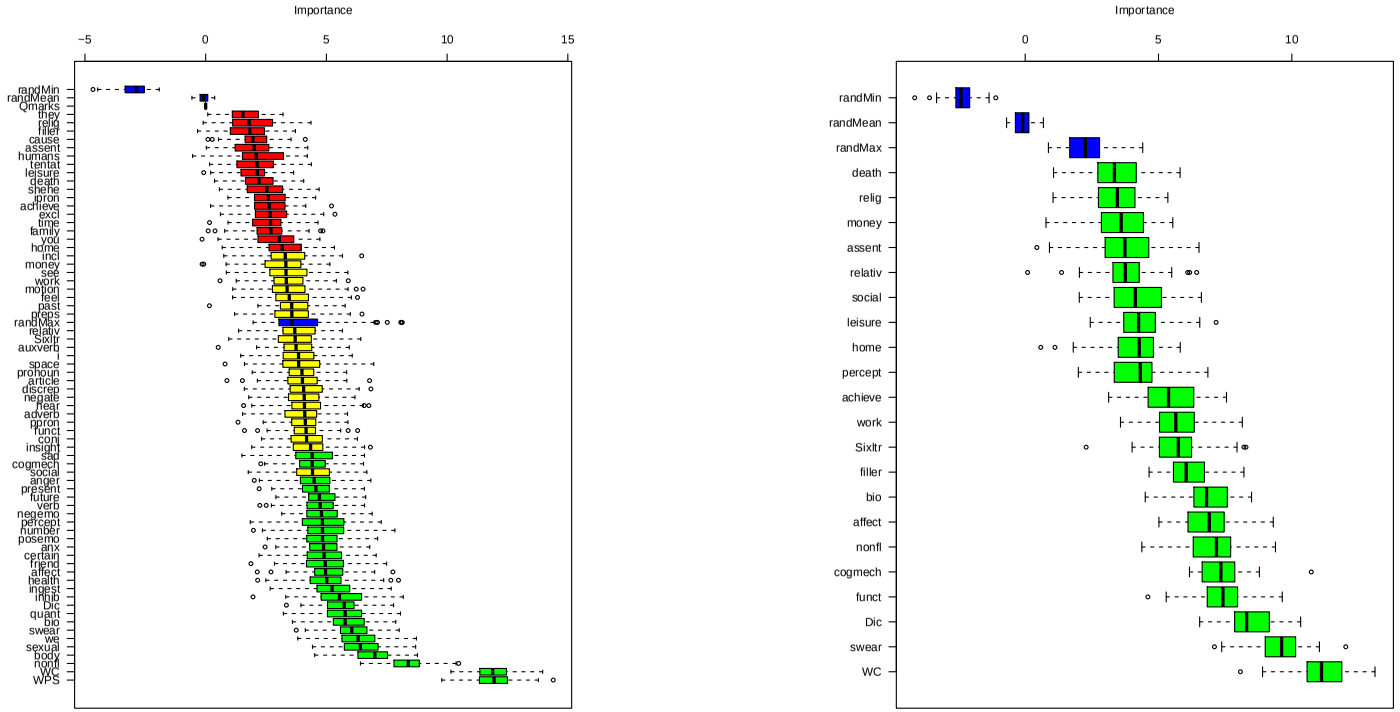
[9] `http://scikit-learn.org`



(a)Boruta Test for Behavioral Features

(b)Boruta Test for Network Features

Figure 4.5: Boruta Feature Importance Test for feature selection to show importance of features (PacSocial experiment). Blue bars indicate importance values for random features, red bars show features which values are lower than maximum random feature values, yellow bars indicate that features are about as relevant as random features and green bars indicate relevant features with higher values than random features. For classification only features with higher values than random features, therefore features plotted as green bars should be used.



(a)Boruta Test for LIWC Features (All Subcategories)

(b)Boruta Test for LIWC Features (Supercategories)

Figure 4.6: Boruta Feature Importance Test for feature selection to show importance of features (PacSocial experiment). Blue bars indicate importance values for random features, red bars show features which values are lower than maximum random feature values, yellow bars indicate that features are about as relevant as random features and green bars indicate relevant features with higher values than random features. For classification only features with higher values than random features, therefore features plotted as green bars should be used.

Classifier Selection

Table 4.5 shows accuracy values for trained classifiers which performed best. Classifiers are explained briefly in section 4.2.2. One can see that random forest classifier (rfc) and gradient boosting classifier (gbc) outperform other trained classifiers by using 10-fold cross validation and only LIWC supercategories as linguistic features. Using linguistic subcategories as features performs slightly worse on average than using only supercategories.

Table 4.5: Comparison of classifiers' performance (PacSocial experiment). The random baseline is given by accuracy values of 0.50. Classifiers are described briefly in section 4.2.2.

	All Features						**Features without LIWC Subcategories**					
	K=10			**K=4**			**K=10**			**K=4**		
Model	**Rec**	**Prec**	**F1**	**Rec**	**Prec**	**F1**	**Rec**	**Prec**	**F1**	**Rec**	**Prec**	**F1**
random	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
svc	0.62	0.62	0.62	0.60	0.60	0.60	0.56	0.56	0.56	0.60	0.59	0.59
linearSvc	0.60	0.60	0.60	0.62	0.61	0.61	0.58	0.58	0.58	0.61	0.59	0.59
rfc	0.57	0.57	0.56	0.57	0.57	0.57	**0.64**	**0.64**	**0.64**	0.61	0.61	0.60
gbc	0.58	0.58	0.58	0.59	0.59	0.59	0.63	0.63	0.63	0.63	0.63	0.63

Fitness of Features for Prediction

For the rfc - the best performing classifier - feature importance values are shown in detail. Table 4.6 shows the importance ranking which sums up to 1 for all features. As one can see in the table, the best 20 features only sum up to approximately 50%.

Figure 4.7 shows the box plots for the top 20 features. Non-susceptible users are represented by yellow boxes and susceptible users by red boxes. One can observe differences of the feature values.

As one can see the best classifier was able to reach a F1 score of about 64%. Although this is significantly more than the random baseline (50%) it is also very distant from a stable classification value.

Interpretation of Results

The boxplot feature values for susceptible and non-susceptible users shown in Figure 4.7 indicate that susceptible users have a higher word count (normalized to the number of messages) than non-susceptible users – indicating they use many words per tweet. Since every tweet is limited to 140 characters that could mean that susceptible users rather consume more characters than non-susceptible users or also that they use shorter but more words. Message count is also higher for susceptible users than for non-susceptible users which indicates that susceptible users are more active. Furthermore, susceptible users show higher values for interaction hubs. That indicates that susceptible users tend to often retweet, reply to or mention users with a high interaction authority score (users with which many other users also interact). A higher lexical variety for non-susceptible

Table 4.6: Importance ranking of the top 20 features (PacSocial experiment) using the random forest classifiers. Values for all features sum up to 1.

Feature	Importance
temporal variety	0.0375
word count	0.0317
work	0.0317
message count	0.0303
biological processes	0.0288
lexical variety	0.0278
interaction hubs	0.0272
question ratio	0.0258
lexical balance	0.0248
social processes	0.0247
affective processes	0.0241
perceptual processes	0.0240
home	0.0240
temporal balance	0.0239
money	0.0233
conversational coverage	0.0233
topical variety	0.0229
conversational balance	0.0228
follow out degree	0.0223
achievement	0.0222
sum	**0.5231**

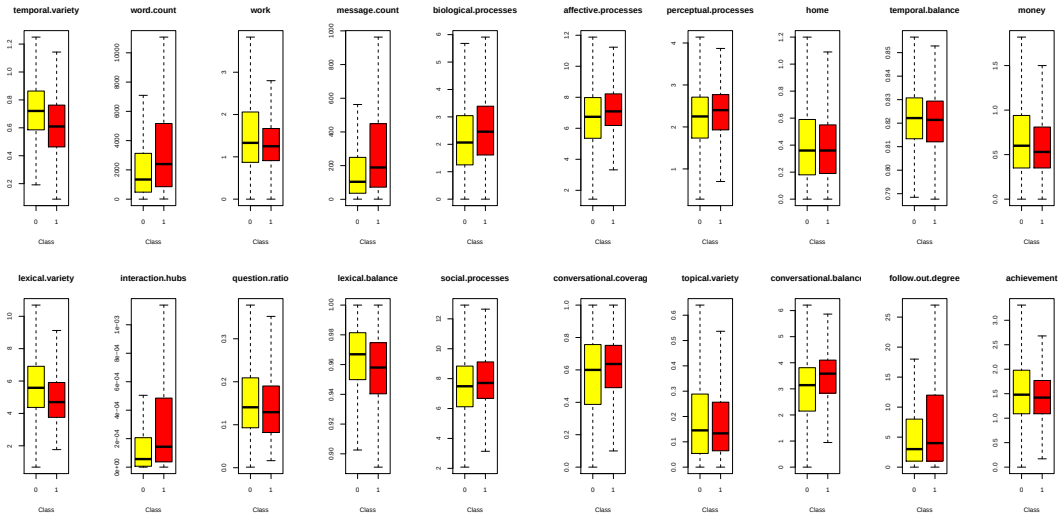


Figure 4.7: Box plots for the top 20 features (PacSocial experiment) according to the feature importance given by the rfc. Yellow boxes (class 0, left) represent non-susceptible users, red boxes (class 1, right) represent susceptible users. Differences between susceptible and non-susceptible users can be observed.

users indicates that they use more distinct words whereas susceptible users seem to use a smaller bag of words. Susceptible users show a higher conversational balance which indicates since the balance is normalized to the number of different users that a high uncertainty exists with whom the user will communicate next, since communicational messages are equally distributed among different users. Nevertheless, it should be noted that the difference between susceptible or non-susceptible samples only differs marginal for many of the top 20 features. Overall, the prediction on which users are susceptible and which not, did perform indifferent for this dataset. That leads to the conclusion that it is hard to predict susceptibility for this target group.

4.2.5 Interpretation of Results comparing the two Experiments

Results for the WebEcology experiment (approx. 71% F1 score) show a higher accuracy for the classification task than results for the PacSocial experiment (approx. 63% F1 score). Comparing boxplots for the top 20 features of each dataset see Figure 4.4 and Figure 4.7 shows that the feature distribution of susceptible and non-susceptible users differ more in the WebEcology experiment than in the PacSocial experiment.

A possible explanation why classifying users performs better on the WebEcology experiment is that this challenge was directed to engage users in conversations or motivate them to follow socialbots. In contrast, the PacSocial study was also directed to create new social links between targeted users. Since the PacSocial socialbots also tried to establish themselves in the existing network, e. g. by following targets and creating tweets directed to targets without link recommendation the susceptibility study was also performed on this dataset, using the same definition for susceptible users. Nevertheless, a smaller fraction of users became susceptible at the PacSocial experiment (13.37%) than at the WebEcology challenge (40.40%). This may indicate that less targets became susceptible due to different socialbot strategies, which could lead to some noise in the non-susceptible user group, since it may therefore include theoretically susceptible users. This can be partially confirmed, since during the classification it was observed that the classifiers' performance tend to depend on which non-susceptible users were randomly chosen for the balanced dataset.

Another possible explanation is that the two socialbot experiments were targeted against two completely different target groups. While for the WebEcology experiment users were chosen by common states, i. e. they were cat lovers, users for the PacSocial experiment seem to be chosen by social relations. The socialbots run on the WebEcology challenge could specifically target users by this additional knowledge and therefore may attract other kinds of users. This indicates that different target groups may react differently to socialbot attacks and show different characteristics for which users become susceptible and which not. Although some of the most important features do not overlap between

the two studies, in both experiments communicative users seem to be more susceptible than others.

4.3 Research Question 3: Social Impact of Socialbots

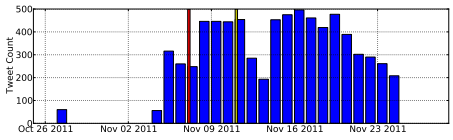
This study is performed according to the third research question – *To what extent and how can socialbots manipulate the link creation behavior of OSN users?*. The study is conducted on the dataset provided by PacSocial described in section 4.1.2. The aim of this study is to explore if and to what extent socialbots can be used to fulfill a special task. In this concrete competition the socialbots were designed to recommend users to each other. Their success rate is measured by analyzing how the shape of the social graph changed through socialbot interaction. In the following section, the design and setup of the second empirical study is described and several measures which allow to assess socialbot impact and success are introduced.

4.3.1 Experimental Setup

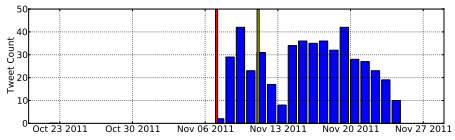
The researchers who run the competition on Twitter found a significant increase of approximately 43% link creation during the experimental period compared to the control period as described in Nanis u. a. (2011). The authors compared the number of newly created links during the control period with the number of newly created links during the experimental period and concluded that the socialbots were very successful in creating new links between users. However, the authors did not further explore if the measured increase is solely caused by the socialbot activities or if other activities may have caused the links.

To investigate this the dataset from the PacSocial challenge is explored in detail. Figure 4.8 shows cumulative number of tweets and recommendation tweets authored by socialbots. One can see that the socialbots did not start tweeting immediately after they were launched at the beginning of the original experimental phase which will be called *experimental phase 1* (exp1) in the following work. Since the experimental phase 1 ended just a few days after the bots started tweeting (marked with a yellow line in the Figures) a second experimental phase is introduced which is called *experimental phase 2* (exp2) in the following. Figure 4.8b shows cumulated recommendation tweets authored by the bots where the red line (first line) indicates the start of the experimental phase 2 and the yellow line (second line) indicates the end of the experimental phase 1.

Figure 4.9 provides an overview of the different observation phases. Above the timeline the original PacSocial experiment is shown with a 33 day control and a 21 day experimental phase called experimental phase 1. Below the timeline the modified experiment



(a)Cumulative Number of tweets authored by socialbots



(b)Cumulative number of recommendation tweets authored by socialbots

Figure 4.8: Tweets and recommendation tweets authored by socialbots where red lines (first) indicate start of experimental phase 2 and yellow lines (second) indicate end of experimental phase 1.

shows the experimental phase 2. This experimental phase lasts 17 days from the 7th November until 24th November. This experimental phase was chosen to start at the same day the socialbots started authoring recommendation tweets which was on the 7th November as previously shown in Figure 4.8b. A second reason to choose this start date was that no follow information is available in the dataset between the 26th October and the 7th November. The last day was chosen by the last available follow information. The control phase did not change for the second experiment.

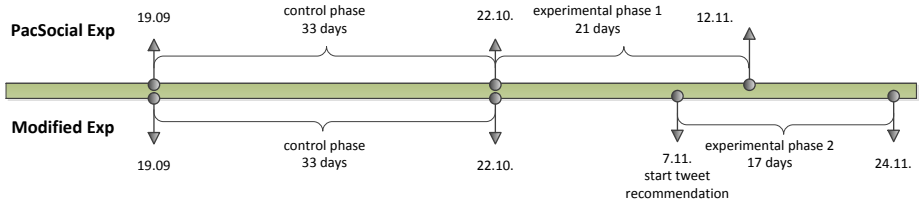


Figure 4.9: Timeline Overview over control and experiment periods. The PacSocial experiment above the timeline shows the original experiment with experimental phase 1 from 22nd October until 12th November. The modified experiment below the timeline shows the second experimental phase lasting from 7th November until 24th November. The control phase stays the same for both experiments.

The aim of the following work is to explore the impact of socialbots in more detail and therefore several success measures, which allow to approximate potential causes of newly created links, are introduced. These measures assess the probability that a newly created link was created caused by a socialbot interaction by assessing the probability of other potential causes (e. g. human interactions or real life happenings) which can or cannot be observed in the data. The introduced measures also help to estimate how many links cannot be explained only by the data and therefore allow to estimate how reliable the results are. If a large proportion of newly created links cannot be explained via socialbot interactions or any other interactions there must be other factors which can explain the link creation behavior. However, those factors may not necessarily be manifested in the data.

Measures for Bot Impact and Success

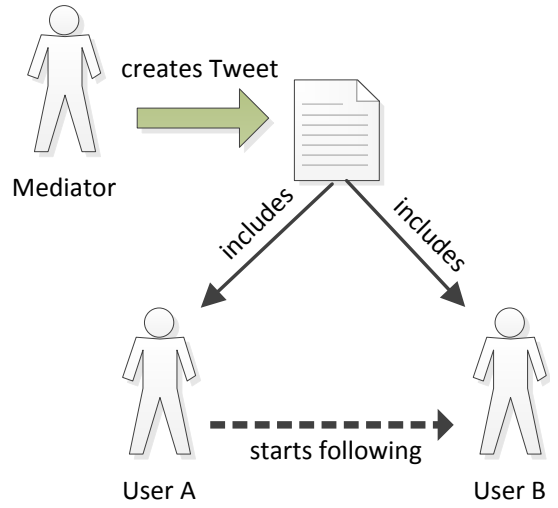
To measure the impact of socialbots while controlling the impact of other confounding variables a set of preceding situations, which may cause the creation of a new link between two human users, is defined. Preceding situations are described on two different dimensions - Recommendation Types and Mediators.

Recommendation Types (RT): Several recommendation types can be identified which may cause a new link between two users, e. g. if two users are mentioned in the same tweet or if two users are connected via a common friend.

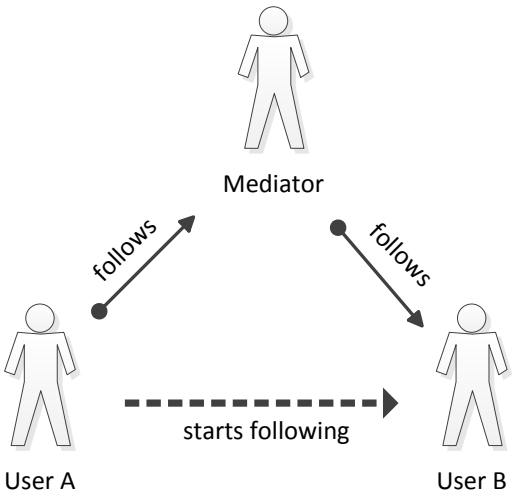
- *RT 1 - Direct User Recommendation via Tweet*: This recommendation type considers links that were created between two users that were previously mentioned together in a tweet (see Figure 4.10a).
- *RT 2 - Indirect Follow Recommendation*: Based on the *triadic closure principle* (Granovetter, 1973) this recommendation type measures newly created links between users that were previously connected by a common mediator (see Figure 4.10b). The triadic closure principle states that if two users A and B have a strong connection with a third user C it is likely that users A and B will also establish a relationship.
- *RT 3 - Indirect User Recommendation via Tweet*: This recommendation type describes the situation when user A creates a follow link to user B after following a mediator who replied to, mentioned or retweeted user B. This recommendation type is visualized in Figure 4.10c.

Mediator: If two users did not have any direct interactions in the past (e. g. via their following or tweeting behavior, see Figure 4.10d) another reason such as a third-party *mediator* could motivate the link creation. A mediator can be a common friend as well as a socialbot who connects two users. However, also real life events or happenings may function as mediating situations.

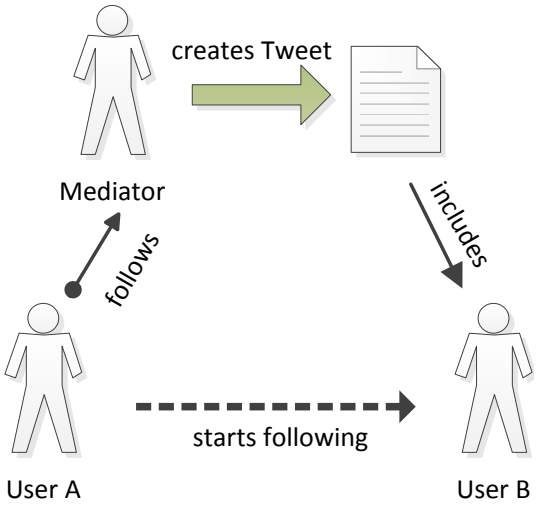
- *Human Mediator:* A link is established between two users after a human mediator performed a specific action. No such preceding socialbot action could be measured related to this link creation. A human mediator can be every user account in the target group, except the socialbots.
- *Socialbot Mediator:* A link is created between two users after a socialbot mediator (but no human mediator) was observed.
- *Human & Socialbot Mediator*: A link is created between two users after a human and a socialbot mediator were observed.
- *No Measurable Mediator:* A link is created between two users but no potential mediator causing the link creation can be identified. This category considers the fact that OSN link creation can exclusively be motivated by real life factors which are not reflected in the dataset captured by the OSNs.



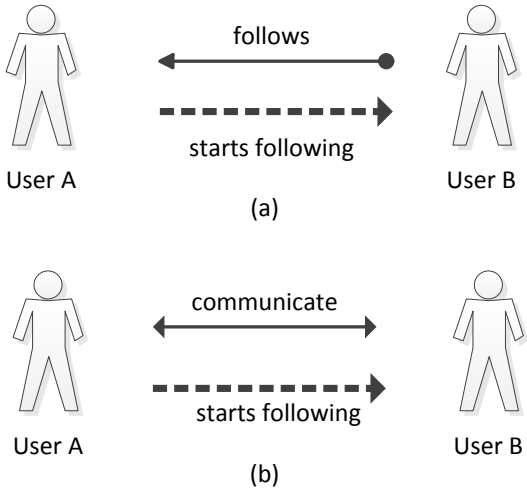
(a) *Recommendation Type 1 (Direct User Recommendation via Tweet)* measures if user A starts following user B after a Mediator mentioned them together in one tweet



(b) *Recommendation Type 2 (Indirect User Recommendation via Follow)* measures if a user A starts following user B after user A followed a Mediator who followed user B



(c) *Recommendation Type 3 (Indirect User Recommendation via Tweet)* measures if user A starts following user B after user A followed a Mediator who replied to, mentioned or retweeted user B



(d) *Direct User Interaction* describes the situation that user A creates a follow link to user B after B followed A previously or after they communicated with each other.

Figure 4.10: This figure shows the different recommendation types and direct user interactions which may cause a new social link between two users. Different recommendation types may involve different mediator types. Direct user interactions do not involve any mediators.

Impact measures are computed by combining recommendation types with different types of mediators. Therefore, introduced measures which capture all potential combinations of recommendation types and involved mediators explain a large variety of possible causes of link creation.

For example one can combine RT 1 and RT 2 (RT 12) with a human mediator. This combination calculates the proportion of newly created links where only a human mediation action can be measured before link creation. The mediation action can look like follows. First, a possible action is that the human mediator mentioned both target users in one previously created tweet (RT 1). Second, the mediator is followed by the source user and follows the target user before the link from source to target user is established (RT 2).

4.3.2 Results and Interpretation

For the PacSocial experiment an average increase of 38.91% in link creation per day was observed in this study while in Nanis u. a. (2011) authors report an average increase of 43% link creation per day. The difference most likely originates from different strategies on handling users that did not exist during the whole phase. Since only newly created links are of interest for this study links which existed at the beginning of the control phase are neglected and only links which existed at the end but not at the beginning are summed. This way, links created and immediately removed afterwards are not considered.

As expected, results vary for different recommendation types and mediator types. Proposed success measures allow to differentiate between

- links which might have been created anyway since the users were already directly or indirectly related before but had no bidirectional connection
- (*human mediated*) links which were most likely caused by human mediators
- (*socialbot mediated*) links, most likely caused by socialbot mediators
- links which were most likely caused by socialbot *and* human mediators (*human & socialbot mediated*)
- links where no direct motivation or mediator can be measured.

Table 4.7 shows link creation in total, link creation with preceding direct user interaction (4.10d) and the calculated difference between those two values for the control phase, the experimental phase 1 and the experimental phase 2. Values are summed over all nine target groups and averaged per day.

Table 4.8 shows the proportion of newly created links per mediator type and recommendation types by applying success measures to the control phase (ctr), experimental phase 1 (exp1) and experimental phase 2 (exp2). The rows of the table correspond to

Table 4.7: The average number of total links which were created on one day during the control phase (*ctr*), during the experimental phase 1 (*exp1*) and during the experimental phase 2 (*exp2*). Further, it shows the average number of links per day, which were created after direct user interaction for each phase. The remaining difference indicates the number of average links per day which might be mediated by something else (e. g. a socialbot).

Link Creation	*ctr*	*exp1*	*exp2*
Total	5.49	7.62	6.47
Direct User Interaction	2.12	2.71	2.77
Basis for Calculations	**3.36**	**4.91**	**3.71**

different recommendation types and combinations of them. Row values for ctr and exp1, and ctr and exp2 sum up to the difference between the total number of link creations minus total number of direct user interaction based link creation, shown in Table 4.7. Values are summed over all nine target groups and averaged per day. The tables show that socialbot mediators used to exploit recommendation strategy 1 (*RT 1*) while humans did not use this recommendation form at all.

Results for Experimental Phase 1

The following results are shown in Table 4.8 for the experimental phase 1. The success of socialbots with their best recommendation strategy (*RT 1*) is only 5.83% although the overall number of link creation increased significantly. Looking at all possible recommendation types (see row *RT123*) shows that during the experimental phase most new links were created without a measurable mediator (54.37%) and a significant part was created after recommendations of human mediators (37.86%). The *human & socialbot mediated* category shows that human and/or socialbots may be responsible for a 4.85% increase in link creation. Finally, only little evidence for the impact of socialbots can be found since only 2.92% of the newly created links were created after socialbot recommendations. The large proportion of mediated links for which no mediator could be observed indicates that for the creation of many social links no potential cause can be identified within the data. This is not surprising since also real world factors may impact the creation of social links and therefore function as mediating events. In summary, results show that the observable impact of socialbots was rather low although the total increase of links during the experimental phase was found to be 38.91% in this work and 43% in Nanis u. a. (2011).

Results for Experimental Phase 2

Interestingly, for experimental phase 2 during which socialbots were far more active than during experimental phase 1 different results can be observed, as shown in Table 4.8. In experimental phase 2, an increase of 17.85% from 5.49 average links per day in the control phase to 6.47 average links per day in the experimental phase 2 was observed. Although the overall number of link creations increases less during the experimental phase 2 than

during the phase 1 a large proportion of new links, which are most likely caused by socialbots (up to 36.51%), can be identified. The proportion of socialbot created links is much higher than for experimental phase 1 and can be explained by the higher intensity of the socialbots during the experimental phase 2.

Results from experimental phase 2 show that indeed a significant proportion of newly created links were caused by socialbots, i. e. many links were created after socialbot mediated recommendations and no other causes could be identified from the data. However, one needs to note that also for a large proportion (up to 47.63%) of newly created links no explanatory causes could be identified from the data which indicates that the link creation process might be influenced by additional factors (e. g. real life happenings) which cannot be acquired from the observational data of OSNs.

Table 4.8: The number of newly created links without the number of links created by *direct user interaction* summed over all target groups averaged per day (see Table 4.7), split by the four mediator types. Values are shown for control phase (ctr), experimental phase 1 (exp1) and experimental phase 2 (exp2). The rows show different recommendation types (RT) as shown in Figure 4.10 and their combinations, where RT 1 shows *Direct User Recommendation via Tweet*, RT 2 describes *Indirect User Recommendation via Follow* and RT 3 describes *Indirect User Recommendation via Tweet*.

Link Creation	*human mediated link creation*						*socialbot mediated link creation*					
Recommendation Type	*ctr*	*%*	*exp1*	*%*	*exp2*	*%*	*ctr*	*%*	*exp1*	*%*	*exp2*	*%*
RT 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29	5.83	1.18	31.73
RT 2	1.46	43.25	2.05	41.75	0.59	15.87	0.00	0.00	0.05	0.98	0.71	19.05
RT 3	0.58	17.12	1.10	22.32	0.12	3.18	0.00	0.00	0.00	0.00	0.00	0.00
RT 12	1.46	43.25	1.86	37.86	0.35	9.53	0.00	0.00	0.14	2.92	1.35	36.51
RT 13	0.58	17.12	1.10	22.32	0.12	3.18	0.00	0.00	0.29	5.83	1.18	31.73
RT 23	1.49	44.14	2.05	41.75	0.59	15.87	0.00	0.00	0.05	0.98	0.71	19.05
RT 123	**1.49**	**44.14**	**1.86**	**37.86**	**0.35**	**9.53**	**0.00**	**0.00**	**0.14**	**2.92**	**1.35**	**36.51**

(a)

Link Creation	*human or socialbot mediated link creation*						*undefined mediated link creation*					
Recommendation Type	*ctr*	*%*	*exp1*	*%*	*exp2*	*%*	*ctr*	*%*	*exp1*	*%*	*exp2*	*%*
RT 1	0.00	0.00	0.00	0.00	0.00	0.00	3.36	100.00	4.62	94.17	2.53	68.24
RT 2	0.00	0.00	0.05	0.98	0.00	0.00	1.91	56.75	2.76	56.31	2.41	65.08
RT 3	0.00	0.00	0.00	0.00	0.00	0.00	2.79	82.88	3.81	77.68	3.59	96.82
RT 12	0.00	0.00	0.24	4.85	0.24	6.34	1.91	56.75	2.67	54.37	1.77	47.63
RT 13	0.00	0.00	0.00	0.00	0.00	0.00	2.79	82.88	3.52	71.85	2.41	65.08
RT 23	0.00	0.00	0.05	0.98	0.00	0.00	1.88	55.86	2.76	56.31	2.41	65.08
RT 123	**0.00**	**0.00**	**0.24**	**4.85**	**0.24**	**6.34**	**1.88**	**55.86**	**2.67**	**54.37**	**1.77**	**47.63**

(b)

Comparing the two Experiments

Comparing findings from both studies shows an increase in link creation in both experiments (see Table 4.7). However, unlike reported in Nanis u. a. (2011), no evidence could be found in the original PacSocial experimental phase 1 that the dramatic increase was caused by socialbots. Figure 4.11a presents the proportion of link creation per mediator

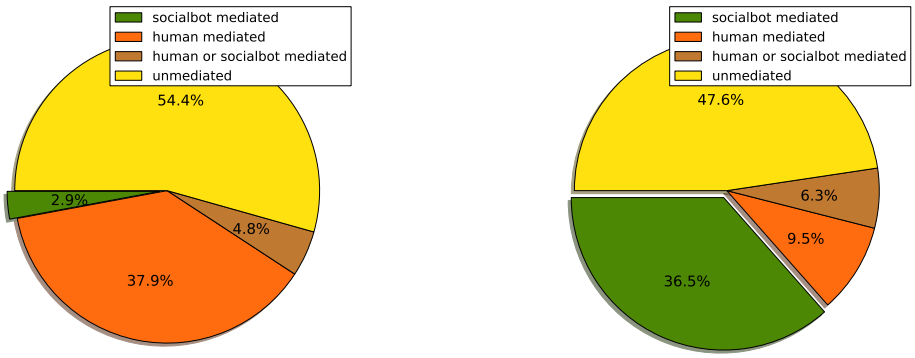
for all three recommendation types and clearly shows that the proportion of links, which were most likely caused by socialbots, is rather small.

For experimental phase 2 the proportion of newly created links, which were most likely caused by socialbots, is significantly higher than for experimental phase 1, as one can see in Figure 4.11b. This can be explained by the fact that the socialbots were more active during the experimental phase 2 than in experimental phase 1. It indicates that the proportion of socialbot caused links can be increased if the attack becomes more intense. However, results also show that hidden factors play an important role and that only around half of newly created links can be explained via recommendation types and mediator types which can be measured on the observational data obtained from Twitter. So it seems that real world events and factors outside the OSN play an important role in the link creation behavior. Though, these factors are hardly contained in the OSN provided data.

Those findings are partly in line with previous studies on predicting social links in OSNs. Other researches also show that predicting social links is a difficult problem since also external factors may impact the link creation behavior of users. For example, in Rowe u. a. (2012) the authors show that, using an extensive set of features the performance of a supervised classification system is pretty low since Matthews Correlation Coefficient (which ranges from -1 to +1 where 0 would be the random baseline) is around 0.1 when using the full dataset. In Backstrom u. Leskovec (2011) the authors show that their approach allows the recommendation of new social links to active Facebook users and achieved a high precision. They show that out of 20 friendships they recommended nearly 40% of them were realized in the near future.

To address the question which socialbot strategy was the most successful of this experiment, the number of socialbot mediated links for each recommendation strategy during the experimental phase 2 is compared. As one can see in Figure 4.12 the recommendation type *direct recommendation via tweet* (see Figure 4.10a) is the most successful, followed by the *indirect recommendation via follow behavior* strategy (Figure 4.10b) whereas the *indirect recommendation via tweet* strategy (Figure 4.10c) could not lead to any link creation at all.

A closer look at the success ratio of socialbot recommendations of type 1 (*RT1*) shows that approximately 4.22% of the recommendation tweets (i. e. tweets in which at least two target users were mentioned) were successful (i. e. the users created a link afterwards). An overall amount of 474 recommendation tweets were created by the socialbots during experimental phase 2. By manually inspecting a sample of those recommendation tweets it could be observed that those tweets usually address one user and recommend one other user (e. g. *@UserA - you would like my #friend @UserB*). However, also tweets, recommending several users to each other at once, are found in the dataset.



(a)Experimental phase 1 (b)Experimental phase 2

Figure 4.11: Proportion of newly created links (without direct preceding user interaction based links) caused by different mediator types and a combination of all three recommender types (RT123) during experimental phases

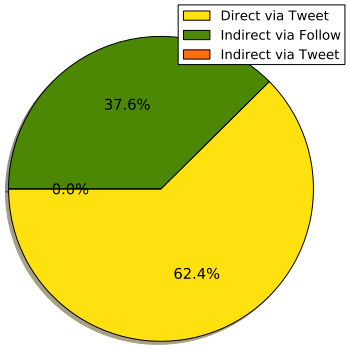


Figure 4.12: Comparison of the success of different recommendation strategies used by socialbots during experimental phase 2.

5 Discussions and Limitations

In the following limitations according to the three research questions identified in section 1.1 are discussed.

Research Question 1 – Taxonomy Proposal

One limitation of the proposed taxonomy is that mainly information about socialbot attacks which took place on Twitter are available. Despite a lot of effort to create a generic categorization system, the proposed taxonomy may be biased towards systems that are similar to Twitter. Also many attacks are only described briefly in the literature since former research mainly concentrated on describing detection of bots and defense mechanisms against bots, and therefore detailed information about attacks may be missing. The attempt to create the taxonomy *generic* enough to cover attacks from different kinds of OSNs while ensuring that it is *specific* enough to be useful (Amoroso, 1994) may impede the categorization of attacks in some cases.

Research Question 2 – Predicting Susceptible Users

One limitation of the empirical study presented in section 4.2 is that the dataset on which the study was conducted only includes tweets from the beginning until the end of the Twitter challenge. Crawling tweets which were published prior to when the challenge was started was not possible due to Twitter API limitations.

For the experiment performed on the PacSocial dataset also semantic features were calculated by using the *Alchemy API*[1]. The API can be used to identify different semantic concepts, keywords and categories. Although the usage of keywords is quite common on Twitter those features did not help to improve performance, therefore they were not included in this work. A possible explanation would be that the dataset consists of too few samples, since many of the identified keywords or concepts were just used once or twice by the users in the dataset. Therefore those feature vectors were very sparse and did not help to improve the predictive model. The author of this thesis did not evaluate the quality of the Alchemy API but Saif u. a. (2012) compared a number of APIs useful to identify semantic meaning from a given text. The best performance results are reported for the Alchemy API with 108 concepts extracted from 500 tweets and an accuracy of approximately 73%.

[1] `http://www.alchemyapi.com/`

Research Question 3 – Potential Impact of Socialbots

One limitation of the empirical study presented in section 4.3 which was conducted on the dataset described in section 4.1.2 is that the dataset only contains information about users within the target groups (i. e. tweets and social relations within a target group). During the control phase only two target groups existed: one consisting of 1,800 and a another one consisting of 900 users. Before the experimental phase started, target groups were reorganized to 9 target groups, each containing 300 users. Information about users' follow behavior during the control phase show that follow links across the target groups also exist and it is unclear how the final target groups were formed. Finally, applying success measures to this dataset only allows excluding the possibility that socialbots try to create new links which would have been created anyway due to the interactions and relations between the targets within their target groups. However, the possibility that socialbots use information about possible indirect relations between users from outside the target groups cannot be excluded since only social graph information within target groups is available for the experimental phase.

6 Conclusions and Outlook

This chapter concludes the work along the three research questions and discusses potential future work.

Research Question 1 – Taxonomy Proposal

This work proposes a taxonomy for socialbot attacks with the aim to provide an overview about the broad and constantly growing field of socialbot attacks on OSNs. The taxonomy was created based on existing examples which show the large variety of socialbot attacks. Nevertheless, it is difficult to create a taxonomy not biased towards a specific OSN such as Twitter, since most information about bot attacks is available for bot attacks which took place on Twitter. The taxonomy needs a level of abstraction detailed enough to be helpful, but also general enough to be applicable for attacks on different OSNs (Amoroso, 1994).

Although creating a categorization system of socialbot attacks is challenging the taxonomy is very helpful for identifying similarities and differences of socialbot attacks. It helps to understand the nature of OSN based socialbot attacks by categorizing them along different dimensions. Especially smaller or relatively new OSNs which never had to deal with socialbot attacks, could benefit from the taxonomy.

In a next step, the taxonomy could serve as a basis to inspect OSNs regarding vulnerabilities and help developing defense mechanisms.

Research Question 2 – Predicting Susceptible Users

This work examined to which extent it is possible to distinguish between users who are susceptible and users who are non-susceptible to socialbot attacks. Various kinds of features were calculated and several classifiers were used to predict which users showed susceptible behavior. The experiment was performed on two datasets what led to (partially) different results. Results for the first dataset indicate that especially users who tend to use Twitter rather for a conversational than an informational purpose become susceptible. The classifier performed well above a random baseline. Classifiers performed a bit worse on the second dataset but also better than a random baseline. Comparing descriptive features for susceptible users for the two datasets shows that the best performing features only partially overlap.

It seems that another set of characteristics have to be identified for different kinds of users to distinguish between susceptible and non-susceptible users.

The bots' behavior may constitute another influencing factor. The WebEcology challenge 4.1.1 clearly showed that the strategies of socialbots (measured by the number of followees and count of interactions with targeted users) may lead to different success ratios regarding the number of susceptible users. Since the investigated datasets resulted from two challenges where socialbots pursued their own goals, this could also be an explanation for the differing results in the experiments.

Overall, this work presents a first important step towards modeling susceptibility of users in OSNs. In a next step, it would be interesting to address exactly those two questions:

- Do different socialbot strategies attract different kind of users?
- Do different kinds of target groups show different features descriptive for becoming susceptible to socialbot attacks.

Research Question 3 – Potential Impact of Socialbots

This work introduced several measures that allow assessing the success of socialbots which aim to create links between users. The utility of these measures is shown within an empirical study based on a Twitter dataset which was crawled before and during a socialbot competition organized by PacSocial. Those measures were defined by combining different Recommendation Types with Mediator Types.

The second experiment shows that socialbots indeed may have influence on link creation between users since for more than 36% of the newly created links (excluding links with preceding direct user interaction) only preceding socialbot interaction was observed.

The results from this empirical study, as well as the results from Nanis u. a. (2011) show that a significant increase of link creation was achieved during the time period where the socialbots were active compared to the time period were no socialbots were active. However, the results of this work show that there is no coercive direct causal relation between the increased social links in the target user group. Further research is required to explore the impact of external factors (e. g. real world happenings) which cannot directly be observed from the OSN data.

A more detailed investigation on whether human mediated link creation is directly influenced by socialbot mediated link creation would be necessary. In a next step *influence models* (Asavathiratham u. a., 2001) may be an option to address this question. Furthermore, Aral u. Walker (2012) investigated which users are rather susceptible and which are rather influential in OSNs. A control group and an experimental group (called *treated group*) randomly chosen from a user's peer network were created to measure different behavior and reducing bias of unobservable factors.

A similar approach could be used in future work to explore *why* users create links. One potential reason could be socialbots. That said setting up such an experiment would not overcome the fact that users could also communicate with each other beyond the artificial group boundaries which would not necessarily be observable but may distort the results. This limitation seems to be not addressed in Aral u. Walker (2012) but using completely unrelated user groups for the experiment may help to overcome this limitation.

Appendix

Table 1: LIWC features from `http://www.liwc.net/descriptiontable1.php`, supercategories are emphasized.

Category	Examples
Linguistic Processes	
Word count	
words/sentence	
Dictionary words	
Words with more than 6 letters	
Total function words	
Total pronouns	I, them, itself
Personal pronouns	I, them, her
1st pers singular	I, me, mine
1st pers plural	We, us, our
2nd person	You, your, thou
3rd pers singular	She, her, him
3rd pers plural	They, their, they'd
Impersonal pronouns	It, it's, those
Articles	A, an, the
Common verbs	Walk, went, see
Auxiliary verbs	Am, will, have
Past tense	Went, ran, had
Present tense	Is, does, hear
Future tense	Will, gonna
Adverbs	Very, really, quickly
Prepositions	To, with, above
Conjunctions	And, but, whereas
Negations	No, not, never
Quantifiers	Few, many, much
Numbers	Second, thousand
Swear words	Damn, piss, fuck
Psychological Processes	
Social processes	Mate, talk, they, child
Family	Daughter, husband, aunt
Friends	Buddy, friend, neighbor
Humans	Adult, baby, boy
Affective processes	Happy, cried, abandon
Positive emotion	Love, nice, sweet
Negative emotion	Hurt, ugly, nasty
Anxiety	Worried, fearful, nervous
Anger	Hate, kill, annoyed
Sadness	Crying, grief, sad
Cognitive processes	cause, know, ought
Insight	think, know, consider
Causation	because, effect, hence
Discrepancy	should, would, could
Tentative	maybe, perhaps, guess
Certainty	always, never
Inhibition	block, constrain, stop
Inclusive	And, with, include
Exclusive	But, without, exclude
Perceptual processes	Observing, heard, feeling
See	View, saw, seen
Hear	Listen, hearing
Feel	Feels, touch
Biological processes	Eat, blood, pain
Body	Cheek, hands, spit
Health	Clinic, flu, pill
Sexual	Horny, love, incest
Ingestion	Dish, eat, pizza
Relativity	Area, bend, exit, stop
Motion	Arrive, car, go
Space	Down, in, thin
Time	End, until, season
Personal Concerns	
Work	Job, majors, xerox
Achievement	Earn, hero, win
Leisure	Cook, chat, movie
Home	Apartment, kitchen, family
Money	Audit, cash, owe
Religion	Altar, church, mosque
Death	Bury, coffin, kill
Spoken categories	
Assent	Agree, OK, yes
Nonfluencies	Er, hm, umm
Fillers	Blah, Imean, youknow

Bibliography

[web 2011] *Socialbots: The End-Game*. http://www.webecologyproject.org/, Feb 2011

[CVE 2012] *Common Vulnerabilities and Exposures, The Standard for Information Sceurity Vulnerability Names*. =http://cve.mitre.org, 06 2012

[Aiello u. a. 2012] Aiello, Luca M. ; Deplano, Martina ; Schifanella, Rossano ; Ruffo, Giancarlo: People Are Strange When You're a Stranger: Impact and Influence of Bots on Social Networks. In: *ICWSM*, 2012

[Alvarez u. Petrovic 2003] Alvarez, Gonzalo ; Petrovic, Slobodan: A new taxonomy of Web attacks suitable for efficient encoding. In: *Computers and Security* 22 (2003), Nr. 5, 435-449. `http://dblp.uni-trier.de/db/journals/compsec/compsec22.html#AlvarezP03`

[Amoroso 1994] Amoroso, Edward G.: *Fundamentals of computer security technology*. Upper Saddle River, NJ, USA : Prentice-Hall, Inc., 1994. – ISBN 0–13–108929–3

[Aral u. Walker 2012] Aral, Sinan ; Walker, Dylan: Identifying Influential and Susceptible Members of Social Networks. In: *Science* 337 (2012), Juli, Nr. 6092, 337–341. `http://dx.doi.org/10.1126/science.1215842`. – DOI 10.1126/science.1215842. – ISSN 1095–9203

[Asavathiratham u. a. 2001] Asavathiratham, C. ; Roy, S. ; Lesieutre, B. ; Verghese, G.: The influence model. In: *Control Systems, IEEE* 21 (2001), dec, Nr. 6, S. 52 –64. `http://dx.doi.org/10.1109/37.969135`. – DOI 10.1109/37.969135. – ISSN 1066–033X

[Backstrom u. Leskovec 2011] Backstrom, Lars ; Leskovec, Jure: Supervised random walks: predicting and recommending links in social networks. In: *Proceedings of the fourth ACM international conference on Web search and data mining*. New York, NY, USA : ACM, 2011 (WSDM '11). – ISBN 978–1–4503–0493–1, 635–644

[Bishop 2006] Bishop, Christopher M.: *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA : Springer-Verlag New York, Inc., 2006. – ISBN 0387310738

[Bishop 1999] Bishop, Matt: Vulnerabilities Analysis. In: *Proceedings of the Second International Symposium on Recent Advances in Intrusion Detection*, McGraw-Hill, 1999

[Boshmaf u. a. 2011] Boshmaf, Yazan ; Muslukhov, Ildar ; Beznosov, Konstantin ; Ripeanu, Matei: The socialbot network: when bots socialize for fame and money. In: *Proceedings of the 27th Annual Computer Security Applications Conference*. New York, NY, USA : ACM, 2011 (ACSAC '11). – ISBN 978–1–4503–0672–0, 93–102

[Boshmaf u. a. 2012] Boshmaf, Yazan ; Muslukhov, Ildar ; Beznosov, Konstantin ; Ripeanu, Matei: Design and Analysis of a Social Botnet. In: *Computer Networks* (2012), Juni. http://dx.doi.org/10.1016/j.comnet.2012.06.006. – DOI 10.1016/j.comnet.2012.06.006. – ISSN 13891286

[Bosma u. a. 2012] Bosma, Maarten ; Meij, Edgar ; Weerkamp, Wouter: A framework for unsupervised spam detection in social networking sites. In: *Proceedings of the 34th European conference on Advances in Information Retrieval*. Berlin, Heidelberg : Springer-Verlag, 2012 (ECIR'12). – ISBN 978–3–642–28996–5, 364–375

[Cao u. a. 2012] Cao, Qiang ; Sirivianos, Michael ; Yang, Xiaowei ; Pregueiro, Tiago: Aiding the detection of fake accounts in large scale social online services. In: *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. Berkeley, CA, USA : USENIX Association, 2012 (NSDI'12), 15–15

[Chu u. a. 2010] Chu, Zi ; Gianvecchio, Steven ; Wang, Haining ; Jajodia, Sushil: Who is tweeting on Twitter: human, bot, or cyborg? In: *Proceedings of the 26th Annual Computer Security Applications Conference*. New York, NY, USA : ACM, 2010 (ACSAC '10). – ISBN 978–1–4503–0133–6, 21–30

[Coburn u. Marra 2008] Coburn, Zack ; Marra, Greg: *Realboy believable twitter bots*. 2008. – http://ca.olin.edu/2008/realboy/results.html

[Danezis u. Mittal 2009] Danezis, George ; Mittal, Prateek: SybilInfer: Detecting Sybil Nodes using Social Networks. In: *NDSS*, 2009

[Douceur 2002] Douceur, John R.: The Sybil Attack. In: *Revised Papers from the First International Workshop on Peer-to-Peer Systems*. London, UK, UK : Springer-Verlag, 2002 (IPTPS '01). – ISBN 3–540–44179–4, 251–260

[Friedman u. a. 2010] Friedman, Jerome H. ; Hastie, Trevor ; Tibshirani, Rob: Regularization Paths for Generalized Linear Models via Coordinate Descent. In: *Journal of Statistical Software* 33 (2010), 2, Nr. 1, 1–22. http://www.jstatsoft.org/v33/i01/paper

[Gao u. a. 2010] Gao, Hongyu ; Hu, Jun ; Wilson, Christo ; Li, Zhichun ; Chen, Yan ; Zhao, Ben Y.: Detecting and characterizing social spam campaigns. In: *Proceedings of the 10th annual conference on Internet measurement*. New York, NY, USA : ACM, 2010 (IMC '10). – ISBN 978–1–4503–0483–2, 35–47

[Ghosh u. a. 2011] Ghosh, Saptarshi ; Korlam, Gautam ; Ganguly, Niloy: Spammers' networks within online social networks: a case-study on Twitter. In: *Proceedings of the 20th international conference companion on World wide web*. New York, NY, USA : ACM, 2011 (WWW '11). – ISBN 978–1–4503–0637–9, 41–42

[Ghosh u. a. 2012] Ghosh, Saptarshi ; Viswanath, Bimal ; Kooti, Farshad ; Sharma, Naveen K. ; Korlam, Gautam ; Benevenuto, Fabricio ; Ganguly, Niloy ; Gummadi, Krishna P.: Understanding and combating link farming in the twitter social network. In: *Proceedings of the 21st international conference on World Wide Web*. New York, NY, USA : ACM, 2012 (WWW '12). – ISBN 978–1–4503–1229–5, 61–70

[Gianvecchio u. a. 2011] Gianvecchio, Steven ; Xie, Mengjun ; Wu, Zhenyu ; Wang, Haining: Humans and bots in internet chat: measurement, analysis, and automated classification. In: *IEEE/ACM Trans. Netw.* 19 (2011), Oktober, Nr. 5, 1557–1571. http://dx.doi.org/10.1109/TNET.2011.2126591. – DOI 10.1109/TNET.2011.2126591. – ISSN 1063–6692

[Granovetter 1973] Granovetter, Mark: The Strength of Weak Ties. In: *The American Journal of Sociology* 78 (1973), May, Nr. 6, 1360-1380. http://links.jstor.org/sici?sici=0002-9602(197305)78:6%253C1360:TSOWT%253E2.0.CO;2-E

[Grier u. a. 2010] Grier, Chris ; Thomas, Kurt ; Paxson, Vern ; Zhang, Michael: @spam: the underground on 140 characters or less. In: *Proceedings of the 17th ACM conference on Computer and communications security*. New York, NY, USA : ACM, 2010 (CCS '10). – ISBN 978–1–4503–0245–6, 27–37

[Hansman u. Hunt 2005] Hansman, Simon ; Hunt, Ray: A taxonomy of network and computer attacks. In: *Computers and Security* 24 (2005), Nr. 1, 31-43. http://dblp.uni-trier.de/db/journals/compsec/compsec24.html#HansmanH05

[Howard 1998] Howard, John D.: *An analysis of security incidents on the Internet 1989-1995*. Pittsburgh, PA, USA, Carnegie Mellon University, Diss., 1998. – UMI Order No. GAX98-02539

[Hwang u. a. 2012] Hwang, Tim ; Pearce, Ian ; Nanis, Max: Socialbots: voices from the fronts. In: *interactions* 19 (2012), März, Nr. 2, 38–45. http://dx.doi.org/10.1145/2090150.2090161. – DOI 10.1145/2090150.2090161. – ISSN 1072–5520

[Igure u. Williams 2008] Igure, V. ; Williams, R.: Taxonomies of attacks and vulnerabilities in computer systems. In: *Communications Surveys Tutorials, IEEE* 10 (2008),

quarter, Nr. 1, S. 6 –19. http://dx.doi.org/10.1109/COMST.2008.4483667. – DOI 10.1109/COMST.2008.4483667. – ISSN 1553–877X

[Jindal u. Liu 2008] Jindal, Nitin ; Liu, Bing: Opinion spam and analysis. In: *Proceedings of the 2008 International Conference on Web Search and Data Mining*. New York, NY, USA : ACM, 2008 (WSDM '08). – ISBN 978–1–59593–927–2, 219–230

[Kleinberg 1999] Kleinberg, Jon M.: Authoritative sources in a hyperlinked environment. In: *J. ACM* 46 (1999), September, Nr. 5, 604–632. http://dx.doi.org/10.1145/324133.324140. – DOI 10.1145/324133.324140. – ISSN 0004–5411

[Krsul 1998] Krsul, Ivan V.: *Software Vulnerability Analysis*. 1998

[Kursa u. Rudnicki 2010] Kursa, Miron B. ; Rudnicki, Witold R.: Feature Selection with the Boruta Package. In: *Journal of Statistical Software* 36 (2010), 9, Nr. 11, 1–13. http://www.jstatsoft.org/v36/i11. – ISSN 1548–7660

[Lee u. a. 2010] Lee, Kyumin ; Caverlee, James ; Webb, Steve: Uncovering social spammers: social honeypots + machine learning. In: *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA : ACM, 2010 (SIGIR '10). – ISBN 978–1–4503–0153–4, 435–442

[Lee u. a. 2011] Lee, Kyumin ; Eoff, Brian D. ; Caverlee, James: Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. In: Adamic, Lada A. (Hrsg.) ; Baeza-Yates, Ricardo A. (Hrsg.) ; Counts, Scott (Hrsg.): *ICWSM*, The AAAI Press, 2011

[Lindqvist u. Jonsson 1997] Lindqvist, Ulf ; Jonsson, Erland: How to Systematically Classify Computer Security Intrusions. In: *Proceedings of the 1997 IEEE Symposium on Security and Privacy*. Washington, DC, USA : IEEE Computer Society, 1997 (SP '97), 154–

[Lough 2001] Lough, Daniel L.: *A Taxonomy Of Computer Attacks With Applications To Wireless Networks*. 2001

[Mervis u. Rosch 1981] Mervis, C B. ; Rosch, E: Categorization of Natural Objects. In: *Annual Review of Psychology* 32 (1981), Nr. 1, 89-115. http://dx.doi.org/10.1146/annurev.ps.32.020181.000513. – DOI 10.1146/annurev.ps.32.020181.000513

[Mevik u. Wehrens 2007] Mevik, Björn-Helge ; Wehrens, Ron: The pls Package: Principal Component and Partial Least Squares Regression in R. In: *Journal of Statistical Software* 18 (2007), 1, Nr. 2, 1–24. http://www.jstatsoft.org/v18/i02/paper

[Mitter u. a. 2013] Mitter, Silvia ; Wagner, Claudia ; Strohmaier, Markus: Understanding the Nature & Impact of Socialbot Attacks in Online Social Networks. In: *WebSci 2013, Proceedings of Web Science 2013*. Paris, France, 2013. – Under Review

[Nanis u. a. 2011] Nanis, Max ; Pearce, Ian ; Hwang, Tim: *PacSocial: Field Test Report*. Nov 2011. – =http://www.pacsocial.com

[Page u. a. 1999] Page, Lawrence ; Brin, Sergey ; Motwani, Rajeev ; Winograd, Terry: The PageRank Citation Ranking: Bringing Order to the Web. / Stanford InfoLab. Version: November 1999. `http://ilpubs.stanford.edu:8090/422/`. Stanford InfoLab, November 1999 (1999-66). – Technical Report. – Previous number = SIDL-WP-1999-0120

[Paulauskas u. Garsva 2006] Paulauskas, N. ; Garsva, E.: Computer System Attack Classification. In: *Electronics and Electrical Engineering* 2 (2006), Nr. 66, 84–87. `http://www.ee.ktu.lt/journal/2006/2/1392-1215-2006-02-66-84.pdf`

[Pennebaker u. a. 2003] Pennebaker, J.W. ; Mehl, M.R. ; Niederhoffer, K.G.: Psychological aspects of natural language use: Our words, our selves. In: *Annual review of psychology* 54 (2003), Nr. 1, S. 547–577

[Quinn u. Bederson 2011] Quinn, Alexander J. ; Bederson, Benjamin B.: Human computation: a survey and taxonomy of a growing field. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA : ACM, 2011 (CHI '11). – ISBN 978–1–4503–0228–9, 1403–1412

[Ridgeway 2005] Ridgeway, Greg: *Generalized boosted models: A guide to the gbm package*. 2005

[Rowe u. a. 2012] Rowe, Matthew ; Stankovic, Milan ; Alani, Harith: Who will follow whom? Exploiting semantics for link prediction in attention-information networks. In: *11th International Semantic Web Conference (ISWC 2012)*, 2012

[Saif u. a. 2012] Saif, Hassan ; He, Yulan ; Alani, Harith: Semantic sentiment analysis of twitter. In: *Proceedings of the 11th international conference on The Semantic Web - Volume Part I*. Berlin, Heidelberg : Springer-Verlag, 2012 (ISWC'12). – ISBN 978–3–642–35175–4, 508–524

[Stein u. a. 2011] Stein, Tao ; Chen, Erdong ; Mangla, Karan: Facebook immune system. In: *Proceedings of the 4th Workshop on Social Network Systems*. New York, NY, USA : ACM, 2011 (SNS '11). – ISBN 978–1–4503–0728–4, 8:1–8:8

[Tausczik u. Pennebaker 2010] Tausczik, Yla R. ; Pennebaker, James W.: The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. In: *Journal of Language and Social Psychology* 29 (2010), März, Nr. 1, 24–54. `http://dx.doi.org/10.1177/0261927X09351676`. – DOI 10.1177/0261927X09351676. – ISSN 1552–6526

[Thomas u. a. 2012] Thomas, Kurt ; Grier, Chris ; Paxson, Vern: Adapting social spam infrastructure for political censorship. In: *Proceedings of the 5th USENIX conference on Large-Scale Exploits and Emergent Threats*. Berkeley, CA, USA : USENIX Association, 2012 (LEET'12), 13–13

[Thomas u. a. 2011] Thomas, Kurt ; Grier, Chris ; Song, Dawn ; Paxson, Vern: Suspended accounts in retrospect: an analysis of twitter spam. In: *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. New York, NY, USA : ACM, 2011 (IMC '11). – ISBN 978–1–4503–1013–0, 243–258

[Tyagi u. G.Aghila 2011] Tyagi, Amit K. ; G.Aghila: Article: A Wide Scale Survey on Botnet. In: *International Journal of Computer Applications* 34 (2011), November, Nr. 9, S. 10–23. – Published by Foundation of Computer Science, New York, USA

[Wagner u. Strohmaier 2010] Wagner, C. ; Strohmaier, M.: The Wisdom in Tweetonomies: Acquiring Latent Conceptual Structures from Social Awareness Streams. In: *Proc. of the Semantic Search 2010 Workshop (SemSearch2010)*, 2010

[Wagner u. a. 2012] Wagner, Claudia ; Mitter, Silvia ; Körner, Christian ; Strohmaier, Markus: When social bots attack: Modeling susceptibility of users in online social networks. In: *Proceedings of the 2nd Workshop on Making Sense of Microposts (MSM'2012)*. Lyon, France, 2012 (held in conjunction with the 21st World Wide Web Conference (WWW'2012))

[Yang u. a. 2011a] Yang, Chao ; Harkreader, Robert ; Gu, Guofei: Die Free or Live Hard? Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers. In: *Proceedings of the 14th International Symposium on Recent Advances in Intrusion Detection (RAID'11)*, 2011

[Yang u. a. 2012] Yang, Chao ; Harkreader, Robert ; Zhang, Jialong ; Shin, Seungwon ; Gu, Guofei: Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In: *Proceedings of the 21st international conference on World Wide Web*. New York, NY, USA : ACM, 2012 (WWW '12). – ISBN 978–1–4503–1229–5, 71–80

[Yang u. a. 2011b] Yang, Zhi ; Wilson, Christo ; Wang, Xiao ; Gao, Tingting ; Zhao, Ben Y. ; Dai, Yafei: Uncovering social network sybils in the wild. In: *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. New York, NY, USA : ACM, 2011 (IMC '11). – ISBN 978–1–4503–1013–0, 259–268

[Yu 2011] Yu, Haifeng: Sybil defenses via social networks: a tutorial and survey. In: *SIGACT News* 42 (2011), Oktober, Nr. 3, 80–101. `http://dx.doi.org/10.1145/2034575.2034593`. – DOI 10.1145/2034575.2034593. – ISSN 0163–5700

[Yu u. a. 2010] Yu, Haifeng ; Gibbons, Phillip B. ; Kaminsky, Michael ; Xiao, Feng: SybilLimit: a near-optimal social network defense against sybil attacks. In: *IEEE/ACM Trans. Netw.* 18 (2010), Juni, Nr. 3, 885–898. http://dx.doi.org/10.1109/TNET.2009.2034047. – DOI 10.1109/TNET.2009.2034047. – ISSN 1063–6692

[Yu u. a. 2008] Yu, Haifeng ; Kaminsky, Michael ; Gibbons, Phillip B. ; Flaxman, Abraham D.: SybilGuard: defending against sybil attacks via social networks. In: *IEEE/ACM Trans. Netw.* 16 (2008), Juni, Nr. 3, 576–589. http://dx.doi.org/10.1109/TNET.2008.923723. – DOI 10.1109/TNET.2008.923723. – ISSN 1063–6692

[Zhao 2003] Zhao, Shanyang: Toward a taxonomy of copresence. In: *Presence: Teleoper. Virtual Environ.* 12 (2003), Oktober, Nr. 5, 445–455. http://dx.doi.org/10.1162/105474603322761261. – DOI 10.1162/105474603322761261. – ISSN 1054–7460

[Zhou u. a. 2003] Zhou, Feng ; Zhuang, Li ; Zhao, Ben Y. ; Huang, Ling ; Joseph, Anthony D. ; Kubiatowicz, John: Approximate object location and spam filtering on peer-to-peer systems. In: *Proceedings of the ACM/IFIP/USENIX 2003 International Conference on Middleware*. New York, NY, USA : Springer-Verlag New York, Inc., 2003 (Middleware '03). – ISBN 3–540–40317–5, 1–20