AN ANALYSIS OF INFORMATION SPREADING AND PRIVACY ISSUES ON SOCIAL NETWORKS

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in Computer Engineering

by
Burcu SAYİN

December 2017
İZMİR

We approve the thesis of **Burcu SAYİN**

Examining Committee Members:

Assoc. Prof. Dr. Orhan DAĞDEVİREN
International Computer Institute, Ege University

Asst. Prof. Dr. Selma TEKİR
Department of Computer Engineering, İzmir Institute of Technology

Asst. Prof. Dr. Serap ŞAHİN
Department of Computer Engineering, İzmir Institute of Technology

25 December 2017

Asst. Prof. Dr. Serap ŞAHİN
Supervisor, Department of Computer Engineering
İzmir Institute of Technology

Assoc. Prof. Dr. Yusuf Murat ERTEN
Head of the Department of
Computer Engineering

Prof. Dr. Aysun SOFUOĞLU
Dean of the Graduate School of
Engineering and Sciences

ACKNOWLEDGMENTS

First of all, I would like to express my gratitude to my supervisor, Asst. Prof. Dr. Serap Şahin, who guided me since the beginning of my education, and encouraged me during this thesis study. I am so grateful to her everlasting support and motivation, which always carried me one step further. It was a big pleasure for me to work with her.

In addition, I would like to extend my warmest appreciation to Assoc. Prof. Dr. Charalampos Z. Patrikakis and Dr. Dimitrios G. Kogias, from Piraeus University of Applied Sciences, in Greece, for their collaboration, continuous support, and extremely helpful comments on my thesis study. I feel so happy and lucky to know them.

What is more, I would like to express my profound thanks to the Council of Higher Education, in Turkey, for supporting and funding a part of my thesis study within the scope of Project-Based Mevlana Exchange Programme, with the project title of "Analysis of Data Aggregation and Privacy Relations on Graph Databases", and project code MEV-2016-057.

Finally, I would like to express my infinite gratitude to my family for their unconditional love, and endless support. It is the most perfect feeling in the world to know that my family is always there.

ABSTRACT

AN ANALYSIS OF INFORMATION SPREADING AND PRIVACY ISSUES ON SOCIAL NETWORKS

With Social Networks (SNs), being populated by a still increasing number of people, who take advantage of the communication and collaboration capabilities that they offer, density of the information, spread over SNs is increasing steadily. Furthermore, the probability of exposure of someone's personal moments to a wider than expected crowd is also increasing. Hence, analyzing the spreading area and privacy level of any information through a SN is an important issue in social network analysis.

By studying the functionalities and characteristics that modern SNs offer, along with the people's habits and common behavior in them, it is easy to understand that several privacy risks may exist, for many of which people may be unaware of. We address this issue, focusing on interactions with posts in a SN, using Facebook as the research domain. As a novelty, we propose an application tool which visualizes the effect of potential privacy risks in Facebook and provides users to control their privacy. The proposed (and simulated) tool allows a Post Owner to observe the spreading area of his/her post, depending on the selected privacy settings of this post. Moreover, it provides preliminary feedback for all the Facebook users that have interacted with this post, to make them aware of the possible privacy changes, aiming to give them a chance to protect the privacy of their interaction on this post by deleting it when such a privacy change takes place.

ÖZET

SOSYAL AĞLARDA BİLGİNİN YAYILIMI VE MAHREMİYET KONULARININ ANALİZİ

Sosyal Ağlar (SA)'ın insanlara sağladığı haberleşme ve işbirliği yapma avantajları ile birlikte, kullanıcı sayısı ve SA üzerinde yayılan bilginin yoğunluğu da giderek artmaktadır. Buna ek olarak, bir kimseye ait özel bilgilerin, paylaşmayı beklediği insan topluluğundan daha büyük bir alanda görülebilme olasılığı da artmaktadır. Bu nedenle, SA üzerinde bilginin yayılımının ve mahremiyetinin analizi SA'nın analizinde önemli bir problemdir.

Modern SA'nın sunduğu servisler, insanların alışkanlıkları ve ortak davranışları ile bir araya geldiğinde, birçok insanın farkında bile olmadığı mahremiyet risklerini oluşturabilir. Tez çalışmasında bu problem hedeflenmiştir. Facebook araştırma alanı olarak kullanılmış ve SA'da kullanıcılar ile gönderiler arasındaki etkileşimlere odaklanılmıştır. Yenilik olarak, Facebook'taki potansiyel mahremiyet risklerinin etkisini görselleştiren ve kullanıcılara mahremiyetlerini kontrol altında tutma olanağı sağlayan bir uygulama aracı sunulmuştur. Önerilen ve benzetimi yapılan bu araç; gönderi sahibine, seçmiş olduğu gizlilik ayarlarına göre gönderisinin yayılım alanını gözlemleme şansı vermektedir. Ek olarak, bu gönderiyi beğenerek ve/veya yorum yaparak iletişime geçmiş olan SA kullanıcılarının mahremiyet konusunda farkındalığını arttıracak ve mahremiyetlerini koruyacak bir çözüm sunulmaktadır. Bunu sağlamak için, gönderinin gizlilik ayarı değiştirildiğinde, bu gönderiyi beğenmiş ve/veya yorum yapmış olan kullanıcılara, önerilen araçtan bir geribildirim iletilebilmekte ve hatta bu tip durumlarda beğenilerinin/yorumlarının otomatik olarak silinmesi sağlanabilmektedir.

To my precious family

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

Table **Page**

LIST OF SYMBOLS

S ..... Susceptible
I ..... Infected
R ..... Removed
ctr ..... Counter value to control interest loss of nodes on spreading the information
rsd ..... Number of Residue Nodes
S_n ..... Number of rounds to terminate information spreading process
m ..... Communication traffic
k ..... Degree of a node
n ..... Population size in a network
L_{pop} ..... List of nodes ordered by popularity levels
Gr_{pop} ..... Groups of nodes per popularity
S_{pop} ..... A sample of g, $g \in Gr_{pop}$
p_F ..... Probability of a node to interact with a post on Friends graph
p_w ..... Probability of a node to withdraw its interaction
p_{FoF} ..... Probability of a node to interact with a post on FoF graph
$len(FG)$ ..... Size of the Friends Graph
$I(F)$ ..... Number of interactions on Friends Graph
$len(FoFG)$ ..... Size of the FoF Graph
$I(FoF)$ ..... Number of interactions on FoF Graph

LIST OF ABBREVIATIONS

SNs . . . . . Social Networks
SA . . . . . Sosyal Ağlar
SI . . . . . Susceptible−Infected Epidemic Model
SIS . . . . . Susceptible−Infected−Susceptible Epidemic Model
SIR . . . . . Susceptible−Infected−Removed Epidemic Model
SPNR . . . . . Susceptible–Positive Infected−Negative Infected−Removed Model
SEIR . . . . . Susceptible−Exposed−Infected−Removed Model
PO . . . . . Post Owner
CO . . . . . Comment Owner
RQ . . . . . Research Question
FG . . . . . Friendship Graph
IG . . . . . Interaction Graph
FoF . . . . . Friends of Friends
GUI . . . . . Graphical User Interface

CHAPTER 1

INTRODUCTION

1.1. Thesis' Aim and Objectives

Social Networks (SNs), which provide communication and/or collaboration opportunities for people, are getting more popular today. In fact, the total number of SN users is estimated as 2.7 billion in 2018 [1]. Although SNs ease the communication, they bring many privacy leakages about the users. In this circumstance, it becomes increasingly difficult for SN users to protect the privacy of their sensitive information. Hence, SN users need some solutions to feel comfortable and safe. To satisfy this requirement, we need a real-time monitoring and analysis of the data (information) on SNs to observe the structure behind the information spreading, and detect existing privacy problems.

While we have this idea in our mind, we started to work on this thesis study by considering the fundamental theorems, models, and mathematical background of information spreading. We realized that there is a strong relation between information spreading and epidemics in literature. Then, we examined the existing models and methods to comprehend the mechanism of the information spreading. Furthermore, we applied the well-known information spreading methods on a SN dataset which we used during this thesis study, and analyzed the test results to see whether they match with the proposed results or not. While working on these concepts, we noticed that the topology of the network, on which the information will spread, has a crucial effect on information spreading. In addition to this, we realized that there are many other factors that affect the spreading of information on SNs today. To comprehend these factors with current approaches to information spreading, and see the improvements according to existing requirements, we also did a literature review of recent years.

After completing this preparatory work, we started to examine the existing privacy risks on SNs stemming from the users' actions and privacy settings. To perform a case study, we first searched for the history of Facebook, which is a famous SN, and its privacy basics. Then, we reviewed the related works that focus on privacy leakages of SNs and

propose some solutions to them. After that, we investigated the Facebook as a user and tried to detect current privacy risks by testing different cases. As a result, we found out two crucial risks which directly affect Facebook users' privacy. The critical point of this thesis study started at this point; we tried to find a user-friendly and valuable solution for the detected privacy problems. All in all, we proposed and simulated a tool which can be used either as an external Facebook application or a central service, served by Facebook.

1.2. Organization of Thesis

The thesis is organized as follows. Chapter 2 covers the all performed studies for the analysis of information spreading. Chapter 3 includes the study of privacy issues on SNs, by especially highlighting the current privacy issues on Facebook, and points out the proposed solution for them. Chapter 4 gives the objective and requirements of the proposed solution, and then provides the experimental works, together with the simulation and applicability details of the solution. Chapter 5 includes the conclusion of the thesis, and provide a detailed explanation of the future work.

CHAPTER 2

AN ANALYSIS OF INFORMATION SPREADING

2.1. Introduction

In recent years, rapidly-growing SNs have started to affect the pattern of information spreading among people. Although there is a strong resemblance between the characteristics of epidemics and the information spread in a population, classical approaches of epidemic models are not enough to model the information spreading on SNs today. Dynamic structure of SNs requires a people-oriented and more adaptable information spreading model to represent the real-world activities. This model should analyze both the characteristics of the networks (i.e. topology) and SN users' actions (i.e. interaction, privacy setting, etc.) in depth. Considering this, the analysis of information spreading should consider some crucial research questions listed below:

- How does a post or personal information spread on SNs?
- Which kind of model reflects the information spreading process on SNs?
- Which method of information spreading is more efficient to use on SNs?
- How the speed of information spreading process is defined and measured on SNs?
- How the topology of SNs affects information spreading?
- Is there a relation between SN users' behavior and the spreading pattern of their personal information?
- Do the privacy preferences of SN users affect the information spreading on SNs or not?
- What are the main factors that directly affect information spreading on SNs?

In fact, to find answer of these questions, first, we should analyze the theoretical background of information spreading, and then we can combine it with the dynamics

of today's SNs, and the privacy effect on information spreading process. Hence, the main objective of this chapter is to comprehend the concept of information spreading in literature.

The rest of this chapter is organized as follows. In Section 2.2, fundamentals of information spreading models are presented including the relation to epidemics. Section 2.3 explains how the topology of network affects information spreading. In Section 2.4, current approaches to information spreading are provided. Section 2.5 includes the evaluation results of information spreading methods on SNAP dataset [2]. In Section 2.6, this chapter is concluded, and the future objectives, where this theoretical background and practical results can be applied are presented.

2.2. Epidemics and Information Spreading

Epidemics can be considered as a serious problem in a society. If someone in a community has a contagious disease, he/she is very likely to infect other people. Assume that one person in a community catches an infection, then, there is a probability that this person infects some of the healthy ones, and make them infected. This contagion effect diminishes in intensity after a while, and the infected ones start to be recovered. Indeed, this is a general example for the spreading process of epidemics.

In literature, spreading processes of the information and epidemics are likened to each other [3]. Epidemics spread for a time and then lose their effect; information is also spread with the same behavior. Size of the area affected by the epidemics depends on population size. It is obvious that the probability of a disease spreading in a crowded area is higher than in a deserted area. Hence, population size is an important determinant in the spreading process of epidemics [4]. The idea is similar in SNs, but instead of considering the whole network as the population, we can think it as the ego-network of the information owner. Hence, in SNs, the size of the ego-network is an important effect in information spreading process. We can say that, a post (information) spreads quickly if the owner of the post has lots of connections.

Mainly, there are many ways to model the epidemics, but the common point of them is the existence of "compartments" [5]. Compartments can be defined as the discrete sets of individuals that constitutes a population. Two most common compartments in literature are Susceptible (S) , and Infected (I) . Susceptible ones are healthy, which means

that they are not infected, yet but have a potential to be an Infected. Infected ones have the disease and they can infect the Susceptible ones. Note that each state (compartment) refers to the number of people in the related group. The epidemic model SI, which has only S and I states, is called as Simple Epidemics [3]. This type of epidemics infects the whole population proportionally to the $\log$ of the population size. Let the population size be n, then epidemics spread with $\log n$.

Figure 2.1 [6] shows the spreading process of SI model. Blue points represent susceptible ones and green points show the infected ones, respectively. Initially, number of susceptible ones is equal to the population size, which is 500. After someone gets infected, the number of susceptible ones starts decreasing, while the number of infected people starts increasing.

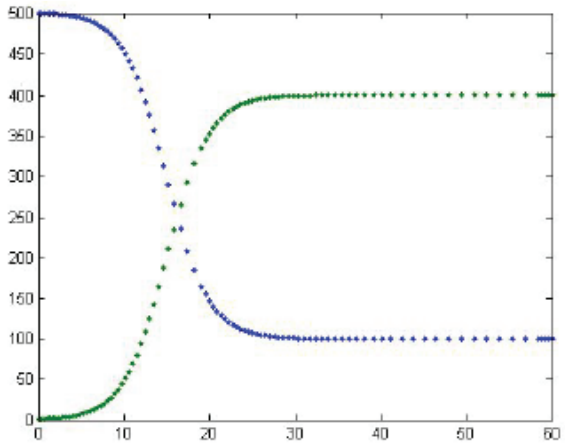

Figure 2.1. Population Size vs Time in SI model
(Blue = Susceptible, Green = Infected)

Yet another epidemic model which can be considered as a two-state model is the SIS (Susceptible−Infected−Susceptible) model [5, 7]. In this model, Susceptible ones can randomly pass to Infected state with an infection rate, which can be considered as a result of the interactions among susceptible and infected ones. In same way, Infected ones can also pass to Susceptible state with a recovery rate, which is defined as the recovering from the disease/infection. Figure 2.2 shows the states and transitions of SIS model.

Apart from two-state epidemic models, there are also many models that includes more states to represent a realistic spreading process. This type of models is referred as Complex Epidemics [3]. The most common state, added by complex epidemic models is Removed (R) state. Removed means the individual recovered from the disease and it

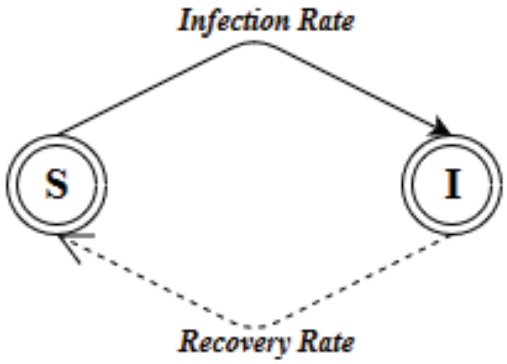

Figure 2.2. States and Transitions of SIS Model

cannot be infected anymore, so it cannot pass to Susceptible state after becoming a Removed. This state can also be considered as vaccinated or immune [5]. One of the oldest and most famous complex epidemic model is the SIR (Susceptible−Infected−Removed) model [5, 8, 9]. SIR model depends on two main assumptions: (i) population has a homogeneous distribution, i.e. individuals interact with each other with an equal probability, and (ii) whole individuals in the population is in Susceptible state initially. Time evolution of a disease in this model can be defined by a threshold theorem, which was proposed by Kermack−McKendrick [8, 9]. This theorem is the ground truth of the state transitions of epidemic models. Considering this, transitions between S, I, and R states are deterministically modeled by the equations 2.1, 2.2, and 2.3.

$$S + I + R = 1. \tag{2.1}$$

$$\frac{dS}{dt} = -SI \tag{2.2}$$

$$\frac{dI}{dt} = +SI - \frac{1}{ctr}(1 - S)I \tag{2.3}$$

Equation 2.2 shows that Susceptible ones will be infected according to the product SI. Equation 2.3 shows an interest loss for the infected individuals in time, so recovery from disease. As it can be seen from the equation 2.3, a counter value (ctr) is added to the formula, where $1/ctr$ shows the probability of interest loss in spreading process of

epidemics. We can find another equation by considering these two equations, and taking a ratio, to specify the infection function $I(s)$ as in equation 2.4.

$$I(S) = \frac{ctr+1}{ctr}(1-S) + \frac{1}{ctr}(logS) \tag{2.4}$$

The function $I(s)$ goes to zero, when the value of S decreases exponentially with the value of ctr, as seen in Equation 2.5 [3, 10].

$$S = e^{-(ctr+1)(1-S)} \tag{2.5}$$

Equation 2.5 shows that if ctr is increased, information reaches a bigger portion of population, but it requires more rounds to complete the spreading process. Hence, ctr provides us to control the termination time of spreading process, and the size of spreading area in a population.

Figure 2.3 shows the states and transitions of SIR model, and Figure 2.4 [6] simulates the spreading process of it.

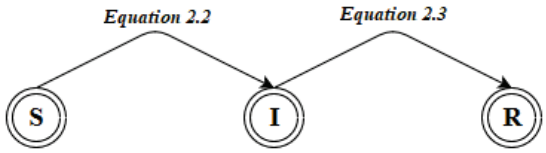

Figure 2.3. States and Transitions of SIR Model

As it can be seen from Figure 2.4 [6], initially all individuals in the population are Susceptible. After an individual gets infected, the number of Infected ones starts increasing, while the number of Susceptible individuals is decreasing. Meanwhile, some of the Infected ones starts to recover from the infectious/disease. Hence, transitions from S to I, and I to R occur according to Equation 2.2 and Equation 2.3, and all individuals become Removed at the end of spreading process.

To comprehend the mechanism behind epidemic models and information spreading in depth, we should examine them in integrity. As spreading a disease in a population, we can utilize the information spreading in a network context to keep all the nodes up-to-date (i.e. each node represents a SN user). Demers et al. [3] define three common methods for performing this propagation update, respectively:

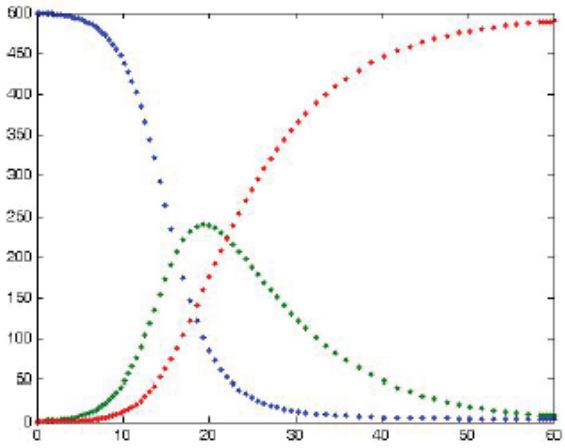

Figure 2.4. Population Size vs Time in SIR model
(Blue = Susceptible, Green = Infected, Red = Removed

1. **Direct mail method:** Each update of a node is mailed/sent to all other nodes in the network. Although this method is efficient, it is not considered reliable because mails (sent updates) can be lost before delivery.

2. **Anti-entropy method:** It is a simple epidemic technique that each node randomly chooses another node to resolve their differences and, then, exchanges the information it holds. There is no difference between Susceptible and Infected nodes in this method. All nodes perform the exchange action in each round and therefore the communication cost increases. This method is less efficient from the direct mail due to the extra processing cost caused by resolving differences. Investigations on the cost of anti-entropy give rise to three update distribution methods: push, pull, and push-pull.

    - According to **push method**, each node randomly chooses another node in each round, if the node has more updated information than the selected node, it pushes the update.
    - **Pull method** does the reverse; having the more updated information, the selected node pulls the up-to-date information.
    - In **push-pull method**, both push and pull methods are applied. Hence, each node randomly chooses another node (neighbor) in each round, if the node has more updated information than the selected node, it pushes the update. Otherwise, the selected node pulls the up-to-date information from this node.

3. The **rumor mongering method** for update propagation adopts complex epidemics mechanisms. This is the only difference from anti-entropy method. Hence, rumor mongering method uses the same update distribution mechanisms (push, pull and push-pull). The method considers all nodes as Susceptible at the beginning. When a node takes any information, it becomes Infected. Only Infected nodes can spread information by selecting random nodes in the network until they lose interest in spreading the information. Hence, there is a fading process, and this process decreases the communication cost. However, this method has some issues as stated below:

    - It is hard to decide when to stop spreading (lose interest/fading process), which is measured by a counter (ctr).
    - To measure the effectiveness of this method, we should specify some expectations for the number of Susceptible, Infected, and Removed nodes. When the spreading process is terminated; the number of Susceptible nodes should be close to 0. This is measured by the count of uninformed or residue nodes (rsd).
    - The speed of information spreading to all network should be maximized. The system should converge to an inactive state (a state that there is no infection, which means spreading is terminated) by the least number of rounds, which is defined by S_n.
    - If ctr value is increased (as mentioned in Equation 2.4), S_n increases. In this case, rsd value decreases, because number of uninformed nodes becomes smaller when the number of rounds is increased.

Demers et al. [3] defines some mechanisms of losing interest in information spreading process, as listed below:

1. **Feedback:** An infected node loses interest only if the recipient (neighbor that is in touch with this node) already has the information.
2. **Blind:** An infected node loses interest with probability $1/ctr$,
3. **Counter:**
    (a) **Counter with feedback:** An infected node loses interest only after communicating with ctr infected nodes.

(b) **Counter with blind:** An infected node loses interest after communicating with ctr nodes.

Considering all the methods and mechanisms mentioned above, we can summarize the crucial points as below:

- The main performance variables of these models are rsd and S_n; rsd should be close to 0, and S_n should be optimized.
- The input variables include ctr and the communication traffic (m). Communication traffic is measured by the number of communication for spreading the up-to-date information in each round [3].
  - The deterministic solutions prove that increasing ctr value with feedback is an effective way of minimizing the values of rsd and S_n.
  - The average value of m for an infected node in each round is formulated as $m = \frac{Total\,information\,update\,count}{number\,of\,nodes}$.
- S_n is proportional to the value of m. Increasing the value of m, on the other hand, decreases rsd according to $rsd = e^{-m}$ [3].

Anti-entropy and rumor mongering methods both use push, pull and push-pull methods as mentioned above. But depending on the nature of the network, the advantages, and disadvantages between push and pull methods vary. For instance; if a network has very frequent multiple information updates simultaneously, then the pull method has advantages in spreading the information very fast because the probability of having information by a node which is chosen randomly to pull the up-to-date information is high. However, if a network has very rare updates, then the pull method creates an unnecessary traffic and in that case, we can prefer push method.

Karp et al. [11] compared the push and pull methods under the same assumptions such as similar update rate, under uniform distribution, and a perfect interconnection without failures. Assume that the network contains n nodes. Push method forwards the information to nodes, and the set of Infected nodes grows exponentially until reaching to half of the network ($n/2$). After this point as shown in Figure 2.5, the set of Susceptible populations shrinks with a constant factor in each round. This factor is about $(1-1/e)$ since the fraction of nodes that do not communicate with any node in a round is approximately $1/e$. Thus, this shrinking phase takes $\theta(\ln n)$ rounds, and the push method sends $\theta(n)$ messages

(information update). In the implementation of pull method; the Infected node has to wait for a connection request to start spreading the information. Therefore, propagation time can be unpredictable for the first round. After the count of Infective nodes reaches to $n/2$ of the population, as shown in Figure 2.6, pull method has an advantage against the push method due to the fraction of Susceptible nodes roughly squares from round to round. The reason behind this is that, in a round starting with $\epsilon.n$ Susceptible nodes, each node has probability $1 - \epsilon$ to get the information, so the probability of staying at susceptible state is ϵ and $\epsilon > 0$. At the end of the round, $\epsilon.n(1 - \epsilon) \approx \epsilon^2.n$ Susceptible nodes will exist. Thus, shrinking phase takes $\theta(\ln \ln(n))$ rounds, and the spreading process include $\theta(n \ln \ln(n))$ messages. "n" factor in the message count comes from the total number of nodes, because each node transmits a message in each round.

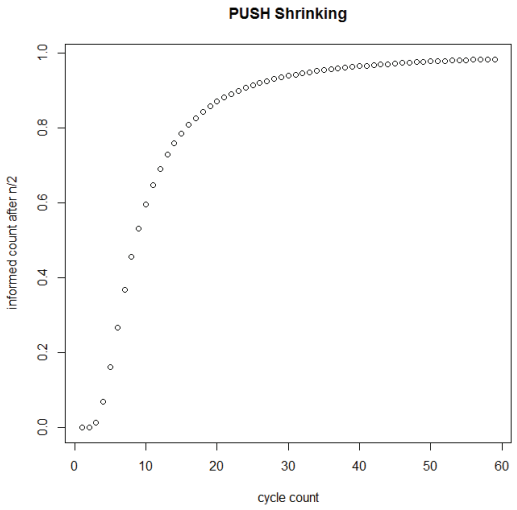

Figure 2.5. PUSH Method Shrinking

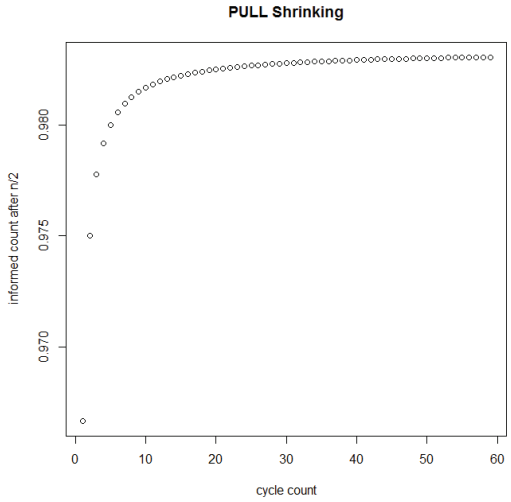

Figure 2.6. PULL Method Shrinking

The goal of Karp et al. [11] was to spread the information through all nodes with minimized number of rounds and update transmission. They impress that mentioned methods are commonly used for lazy transmission of updates. They investigate that this large communication overhead is coming from the nature of epidemic algorithms, and it can be reduced significantly when the information is sent in both directions (i.e. push-pull method). Note that push method works efficiently until $n/2$ of the population becomes Informed, and pull method works efficiently only after $n/2$ becomes informed. Thus, by using these two methods at the same time, we can get a more efficient solution. With a simple push-pull algorithm, spreading any information to whole network requires $S_n = O(\ln n)$ rounds and $O(n \ln \ln(n))$ transmissions.

As an example to the models that use push-pull method and includes a ctr approach in spreading process, we can think of the random phone call model [11]. This model includes randomized communication of n players in parallel rounds, so the robustness of information spreading model increases. In each round, a node (u) randomly picks another node (v). Then, u makes a phone call to v and they exchange the information they hold. In this model, information can be exchanged in both directions (push-pull) in a round t. Hence, the number of transmissions can be reduced significantly by using a simple push-pull algorithm [11]. This model informs all nodes in time with the maximum number of rounds $S_{n_{max}} = \log_3 n + O(\ln \ln(n))$, using $O(n \ln \ln(n))$ messages.

A critical point in this model is determining the optimal termination time. Additionally, it is very sensitive to any error among nodes, which affects the spreading process of information. To improve this model, Karp et al. [11] devise a distribution termination scheme, which is called as median-counter algorithm. According to this algorithm, information is defined as r, and there are four types of states; A, B, C and D. Nodes that has not taken any information yet belongs to state A. For nodes in state B, algorithm holds a counter value, which is shown as $ctr(v, r)$, where v denotes the node. Each time a node in state B takes a new information r, its counter is increased. If $ctr(v, r) = x$, then node v is in B-x state. If x is equal to the maximum counter value, state is changed to C. If a node v in state A receives r only from nodes in state B, then its state is switched to state B-1. If a node in state A receives r from a node in state C, then its state is switched to C. For nodes in state C; every node stays in this phase for at most $ctr_{max} = O(\ln \ln n)$ rounds, and then switches to state D which terminates the information spreading. Hence, the median-counter algorithm uses only $S_n = O(\ln n)$ rounds with $O(\ln \ln n)$ message transmission. How the rsd value is affected and why median-counter algorithm is used are explained in the article [11] thoroughly.

All the epidemic models and spreading methods provided in this section can be considered as an overall background of information spreading. However, the effectiveness of the applied spreading model also depends on the topology of the network, on which the information will be spread. Hence, a discussion about the topology effect on the information spreading process is provided in the following section.

2.3. Topology Effect on Information Spreading

Karp et al. [11] point out an important matter: whether the topology of a network affects information spreading or not. Many studies in literature proposes that the topology of the network is an important parameter to consider while analyzing the information spreading process [12, 13, 14, 15, 16, 17].

As mentioned in previous section, we can evaluate information spreading as a network context to keep all the nodes updated [4]. Mihail et al. [13] propose a model, which considers preferential attachment to create a scale-free network. This model is so important for our study because social networks generally have scale-free property. The model focuses on two important criteria: congestion and conductance. Congestion is a kind of traffic that occurs in a network through some edges that have a bridge property. If congestion is high in a network, conveyance between nodes becomes slow. Conductance mostly refers to the conveyance success rate of a network [18]. If the conductance value of a network is high, then conveyance success rate is also high. Furthermore, it is proposed that congestion is an important effect for the performance.

Mihail et al. [13] mainly focus on two points: (i) "in constant-degree trees congestion grows as n^2 with nodes", (ii) "in constant-degree expanders this growth is close to $n \log n$",which is theoretically minimum. Their model of growth depends on the preferential attachment as mentioned above. According to it, the probability of a node being selected is proportional to its degree (k), so nodes which have bigger degree have bigger probability to be selected. Hence, a scale-free network comes out. They show that, "for $k \geq 2$, almost all scale-free graphs in this model have constant conductance." [13].

Similar to the inferences of Mihail et. al. [13], Lattanzi et al. [12] proposed that if the conductance of a graph is high enough, then information spreading is fast. They show that if an n-node connected graph G has conductance ϕ, then information spreading successfully broadcasts a message within $S_n = O(\frac{\log^4 n}{\phi(G)^6})$ steps, with high probability, using the push-pull method. Furthermore, they showed a relationship between the graph sparsification and information spreading. Spielman and Teng's spectral sparsification procedure [14] is given as the reference for this study. According to it; there is a graph G and sampled graph (ST) of G as $ST(G) \subseteq G$. $ST(G)$ is constructed with same vertices by doing random choices from the edges with probability $p_{u,v} \coloneqq \min\left\{1, \frac{\delta}{\min\left\{deg(u), deg(v)\right\}}\right\}$

where $deg(u)$ denotes the degree of a node u and, $\delta = \Theta\Big(\frac{\log^2 n}{\phi^4}\Big)$. Then, adjacency matrix of $ST(G)$ is found and eigenvalues are calculated. Finally, eigenvalue spectrum of $ST(G)$ comes out which is a good approximation of graph G. It is stated that information spreading stochastically dominates this ST.

As an extension to studies related to graph conductance, Lattanzi et al. [15] observed the convergence time of push-pull method on graphs, which have conductance ϕ. Their result shows that any information spreads within $S_n = \bar{O}(\phi^{-1}.\log n)$ steps, where n is the number of nodes and the notation $\bar{O}(\ldots)$ hides a $polylog\phi^{-1}$ factor, with high probability by using push-pull strategy. They state that this result is almost tight because of holding a graph of n nodes, which has conductance ϕ with diameter $\Omega(\phi^{-1}.\log n)$.

While some researchers focus on conductance on information spreading scope, some of them consider the effect of "weak conductance". Hillel et al. [16, 17] explained the term as follows: "Weak conductance, $\phi_c(G)$, of a graph G, measures connectivity among subsets of nodes in the graph, whose size depend on the parameter $c \geq 1$ ". The most important reason of considering weak conductance in the concept of information spreading is that, we can divide the original graph into its subsets, and then calculate the spreading area properly by measuring the conductance of these subsets. Hence, spreading process of any information becomes well-controlled.

Hillel et al. [16] provide an algorithm, which overcomes the bottlenecks between the nodes of a graph and propose a fast information spreading process among all nodes. They use the idea of weak conductance, that is independent from the conductance as a tool in their study. They state that a graph may have small conductance, but large weak conductance. With the condition of $c \geq 1$ and $\delta \in (0, \frac{1}{3c})$, their algorithm spreads information to all nodes of a graph in $S_n = O\Big(c\Big(\frac{\log n + \log \delta^{-1}}{\phi_c(G)} + c\Big)\Big)$ rounds, with at least the probability of $(1 - 3c\delta)$.

As we can see from the mentioned studies, topology of a network directly affects information spreading. Hence, we should first analyze the structure of network that we would like to work on, and then we should determine the algorithm/method we will implement for the information spreading process.

2.4. Current Approaches to Information Spreading

Although most current studies consider the SIR model as a baseline and modify it according to today's requirements, such as popularity of the information source, content of the information, etc., some of them also propose new approaches with cascades. Information cascades allow us to predict how well the information will spread [19]. In this section, we will firstly introduce the studies, which focus on the modified version of SIR model, and then demonstrate an information spreading model based on cascades.

Bao et al. [20] revises SIR model and divides the Infected state into two: (i) Positive Infected (nodes that have been infected, and they support the information) and (ii) Negative Infected (nodes that have been infected, but they oppose the information). Their model is called as Susceptible−Positive Infected−Negative Infected−Removed (SPNR). According to this model, when a Susceptible node takes the information from a Positive/Negative Infected one, it changes its state to Positive/Negative Infected, with some probability (probabilistic explanations can be found in the original paper [20]). If a Positive Infected node takes the information from a Negative Infected one, it either becomes a Negative Infected or keeps its own state with some probability. When a Negative Infected node takes the information from a Positive Infected, it either becomes a Positive Infected or remains in the same state with some probability. If a Positive/Negative Infected one meets a Removed node, it becomes Removed with some probability. They define the case of turning into removed state with a spreading threshold [20].

Serrano et al. [21] considers an agent-based information spreading model, which based on four states: (i) Neutral (initial state), (ii) Infected (believe the information), (iii) Vaccinated (believe the anti-information before being infected) and (iv) Cured (believe the anti-information after being infected). According to this model, all users are initially neutral. Then, they assign some of them as Infected. Infected ones start to infect their neutral neighbors with a given probability. To simulate cured or vaccinated ones, they define a time, as delay. At that time, a randomly selected Infected user starts to spread anti-information, which says the opposite of the original information in the network. Hence, they try to cure or vaccinate their neighbors with a probability of accepting or denying ($probAcceptDeny$) [21]. Finally, Cured and Vaccinated ones try to cure or vaccinate their neighbors with the value of $probAcceptDeny$.

Cordasco et al. [22] evaluate the Infected state of the SIR model with a different approach. They propose that a user may not immediately start spreading just after it

becomes an Infected; they define a new state for this situation: "Aware". They claim that there should be a threshold that controls the transition from being aware to start spreading. This model resembles the Susceptible−Exposed−Infected−Removed (SEIR) epidemic model [23], which differs from SIR model with the additional "Exposed" state. This state contains people who had contact with an Infected user but have not yet started to infect other people. Similarly, Cordasco et al. [22] propose three states: (i) Ignorant, (ii) Aware and (iii) Spreading. As usual, all users are Ignorant (Susceptible) initially. When an Ignorant node takes information from a Spreader, it becomes Aware. To be a Spreading one, any Aware user should take the information from more than a pre-defined number (threshold value) of Spreading users. This model has no state for Removed, but they define a termination rule in the original paper [22].

Sumith et. al. [24] claims that the assumptions made in SIR model, which was mentioned in Section 2.2 fail in real world. They propose that the whole population do not mix with each other equally, so the distribution is not homogeneous. Moreover, all individuals are not Susceptible initially, and most of them will restrain themselves from interactions. With this approach, authors proposed R_nSIR model, as an extension to SIR model. Hence, they added a new state to SIR model, which is called as R_n. R_n state represents nodes, who restrain themselves from any interaction with other nodes. Authors claims that, in the context of social networks, individuals who are new to network, do not interact well. Hence, they can be accepted as restrained. That is why, they assume that all nodes are in R_n state initially. Transitions of R_n to S, S to I, and I to R, are defined with parameters α, β, and γ, respectively. The parameter α defines the interaction rate of nodes, which is calculated by the number of node's activities on the network. Parameter β defines the influence strength of neighbor node that tries to infect any node in network. Parameter γ defines the recovery rate. Experimental results and the rate change equations of this study can be found in the original article [24].

Tong et al. [25] describes an information cascade model in SNs. First, they provide an extensive study on cascade scales, the scope of the cascade subgraphs, and topological attribute of spread tree. Then, based on the evaluation results, they analyze the spread of the user's decisions for city-wide activities. Decisions include "want to take part in the activity" and "be interested in the activity". This study introduces three mechanisms to use for making a decision:

- **Equal probability:** A node has an equal probability to make any of two decisions.

- **Similarity of nodes:** Similarity of nodes is the criteria to make a decision for any node.
- **Popularity of nodes:** Popularity of nodes affects their decision.

Experiment results of Tong et al. [25] show that popularity of nodes is an important criterion for information spreading. Hence, this study confirms that the information spreading models should also consider user-specific parameters to adopt today's SNs.

Overall, the main aspect of the current approaches for modeling the information spreading on SNs is to propose a realistic model that matches with the complex and dynamic mechanism of human behavior. Hence, researchers try to adopt their models with new parameters, such as popularity of nodes, similarity between them, etc.

2.5. Evaluation of Information Spreading Models on Social Networks

2.5.1. SNAP Dataset

We performed the experiments on SNAP Facebook dataset [2]. This dataset contains real data from Facebook, including users, profile features, friendship relations, etc. However, all these data are kept as anonymized to protect user's privacy. The dataset contains 4039 real Facebook users and 88234 relations between them. We used NetworkX package [26] to create a graph from this dataset to represent a SN. Hence, the resultant social graph includes 4039 nodes and 88234 edges as depicted in Figure 2.7. Each node in this graph represents a real Facebook user, and each edge represents a friendship relation on Facebook. Nodes and edges may have features, such as gender, age, education, political view etc. According to an example in SNAP Facebook Dataset Webpage [2], if a user supports "Democratic Party", and specify this in his/her profile, then this dataset keeps a feature for this user, such as "political=anonymized feature 1". Hence, instead of showing the real information, it keeps some anonymized values for each feature.

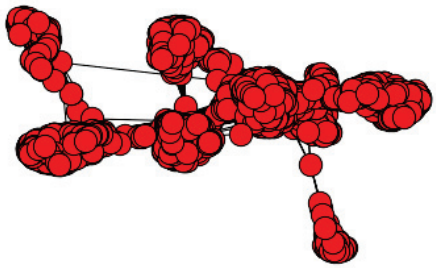

Figure 2.7. Snap Facebook Dataset Representation
Drawn with Python language, using NetworkX Package

2.5.2. Topology of SNAP Dataset

As highlighted in Section 2.3, topology of a graph (network) that we want to work on is so important to analyze information spreading. That is why, we examined the topology of the graph, created from SNAP Facebook Dataset. It can be seen from Figure 2.7 that some nodes have limited number of neighbors (small degree) while others have many. This is the main property of a common scale-free network as SNs. In fact, the distribution of edges refers to a power law distribution. Hence, in the implementation of an information spreading algorithm, the probability of a node being selected is proportional to its degree (k), so nodes which have bigger degree have bigger probability to be selected. This property was also described in Section 2.3 as following: if degree of nodes in a graph is equal to or more than 2, then almost all scale-free networks have constant conductance [13].

2.5.3. Experimental Study on Information Spreading Methods

To observe the spreading of any information through whole dataset, we used SIR model. For the implementation, we made an assumption as follows: an infected node becomes removed only if no Susceptible node remains in the network. The reason behind this assumption is that we wanted to spread information in a fast way, so we should keep the number of Infected nodes as much as possible during the spreading process. Hence, we tuned the recovery idea in the SIR model and combined it with the adaptation of update

distribution methods (push, pull, and push-pull), which were proposed by Karp et al. [11].

Three states exist in the implemented model: (i) Susceptible, (ii) Infected, and (iii) Removed. All nodes are Susceptible initially. Then, a random node is selected, and the information is given to that node, so it becomes an Infected, and starts to spread. Each Susceptible node, that takes the information from an Infected node, becomes an Infected, and spreads the information until the whole graph becomes Informed. In the end, all nodes become Removed, and the spreading process terminates. The important point is that, an Infected node can only infect their neighbors. This is because we assumed that a user can only get information from their friends on Facebook. This assumption may cause this algorithm never to terminate, if the graph is not a connected graph. However, this is not the case for our dataset, because it is a connected graph. What is more, there is no requirement for a ctr value to define removal times in this model, because the spreading process terminates when all nodes in graph (all users in the network) receive the information. When we adapt the push, pull, and push-pull methods to this scenario, we created three different algorithms to implement on the dataset. Algorithms for each method are explained below in detail.

1. **Push Method:** As shown in PUSH Procedure, two lists are kept, respectively: (i) $susceptibleNodeList$, and (ii) $infectedNodeList$. $susceptibleNodeList$ contains all nodes in the graph and $infectedNodeList$ is empty, initially. First round starts after a node is randomly selected from $susceptibleNodeList$, and the information is given to that node. The selected node becomes Infected, so it is removed from $susceptibleNodeList$, and added to $infectedNodeList$. One round is completed when all nodes in $infectedNodeList$ randomly select a node among its neighbor nodes, and try to push the information (no need to push if the selected neighbor already has the information). Each new Infected node is removed from $susceptibleNodeList$, and added to $infectedNodeList$, so list sizes changes dynamically. Push method runs until the whole nodes in the graph become Infected.

2. **Pull method:** As shown in PULL Procedure, only one list for Susceptible nodes is enough for this method. This time, instead of Infected ones, Susceptible nodes randomly selects a neighbor, and try to pull the information. All nodes in the network are Susceptible at the beginning, so $susceptibleNodeList$ holds the whole network initially. After randomly selecting a node and giving the information, that node becomes Infected, and it is removed from $susceptibleNodeList$. Then, first round

starts. One round is completed when all nodes in $susceptibleNodeList$ randomly selects a neighbor node in network, and tries to pull the information (if the selected neighbor does not have the information, no need for the pull operation). Each new Infected node is removed from the $susceptibleNodeList$. Pull method terminates when whole nodes in the network become Infected.

3. **Push-pull method:** PUSH-PULL Procedure depicts the steps of this method. Whole network is kept in a list ($nodeList$). First round starts after a node is randomly selected and the information is given to that node. One round is completed when all nodes in the list randomly choose a neighbor node and either push/pull the information, or does not perform any operation. If the current node has the information, but randomly selected neighbor node does not have, current node pushes the information to this neighbor. If the current node does not have the information, but the randomly selected node has, it pulls the information from this neighbor. If both the current and randomly selected nodes have the information or neither of them have it, no operation is needed. Push-pull method runs until the whole nodes in the network become Infected.

---

procedure PUSH($val, network$) ▷ val is any information, and network is a graph
  $infectedNodeList \leftarrow \varnothing$
  $susceptibleNodeList \leftarrow network.nodes()$
  $randomNode \leftarrow random(network.nodes())$ ▷ a node is randomly selected
  $randomNode.val \leftarrow val$
  $susceptibleNodeList.remove(randomNode)$
  $infectedNodeList.add(randomNode)$
  $numberOfRounds \leftarrow 0$
  while $infectedNodeList.size() \neq network.size()$ **do** ▷ Algorithm is terminated when whole nodes become Infected
    for Each node in $infectedNodeList$ **do**
      Randomly select a $neighborNode$
      if $neighborNode.val = \varnothing$ **then**
        $neighborNode.val \leftarrow val$
        $susceptibleNodeList.remove(neighborNode)$
        $infectedNodeList.add(neighborNode)$
      end if
    end for
    $numberOfRounds \leftarrow numberOfRounds + 1$
  end while
  return $numberOfRounds$
end procedure

```
1: procedure PULL(val, network)      ▷ val is any information, and network is a graph
2:     susceptibleNodeList ← network.nodes()
3:     randomNode ← random(network.nodes())      ▷ a node is randomly selected
4:     randomNode.val ← val
5:     susceptibleNodeList.remove(randomNode)
6:     numberOfRounds ← 0
7:     infectedNodeCount ← 1
8:     while infectedNodeCount ≠ network.size() do      ▷ Algorithm is terminated
   when whole nodes become Infected
9:         currentInfectedNodes ← ∅
10:        for Each node in susceptibleNodeList do
11:            Randomly select a neighborNode
12:            if neighborNode.val = val then
13:                node.val ← neighborNode.val
14:                currentInfectedNodes.add(node)
15:                infectedNodeCount ← infectedNodeCount + 1
16:            end if
17:        end for
18:        for Each node in currentInfectedNodes do
19:            susceptibleNodeList.remove(node)
20:        end for
21:        numberOfRounds ← numberOfRounds + 1
22:    end while
23:    return numberOfRounds
24: end procedure
```

```
1: procedure PUSH-PULL(val, network)      ▷ val is any information, and network is a
   graph
2:     nodeList ← network.nodes()
3:     randomNode ← random(network.nodes())      ▷ a node is randomly selected
4:     randomNode.val ← val
5:     numberOfRounds ← 0
6:     infectedNodeCount ← 1
7:     while infectedNodeCount ≠ network.size() do      ▷ Algorithm is terminated
   when whole nodes become Infected
8:         for Each node in nodeList do
9:             Randomly select a neighborNode
10:            if (node.val = val)and(neighborNode.val = ∅) then
11:                neighbor.val ← node.val
12:                infectedNodeCount ← infectedNodeCount + 1
13:            else if (node.val = ∅)and(neighborNode.val = val) then
14:                node.val ← neighbor.val
15:                infectedNodeCount ← infectedNodeCount + 1
16:            end if
17:        end for
18:        numberOfRounds ← numberOfRounds + 1
19:    end while
20:    return numberOfRounds
21: end procedure
```

Figure 2.8, 2.9, and 2.10 shows the results of the implementation of above three algorithms on the dataset. Figure 2.8 belongs to the push algorithm, and shows the number of Infected node in each round. The algorithm takes almost 3500 rounds to terminate, because after half of the network (red point) becomes informed about the information, it is hard for an infected node to find a neighbor node, which does not have the information. It can be seen from the Figure 2.8 that it takes just few rounds to inform half of the network, but then, the number of rounds exponentially increase. Figure 2.9 shows that, in pull method, after $n/2$ of the population (red point) becomes informed, whole network is informed in a fast way because the probability of a randomly chosen neighbor node to be an infected is high. Hence, pull algorithm is completed in almost 35 rounds. This result supports the study of Karp et al. [11], which propose that if the population size is n ($n = 4039$ for our case), pull method spreads the information faster than push method after $n/2$ of the population being informed (for our case, after the infected count becomes approximately 2019). Figures 2.8 and 2.9 shows the differences.

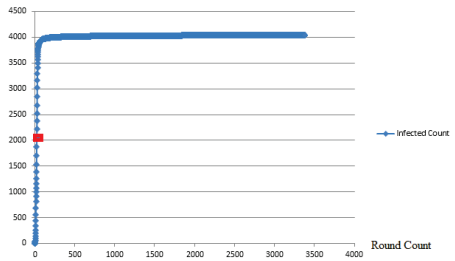

Figure 2.8. PUSH Method

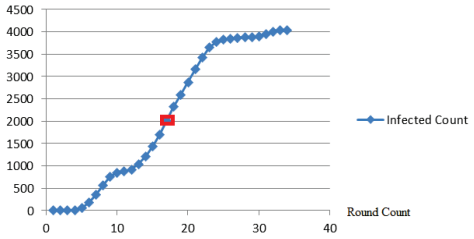

Figure 2.9. PULL Method

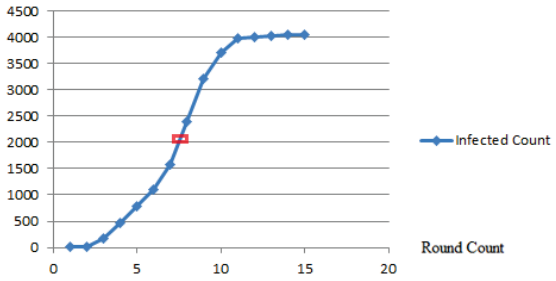

Figure 2.10. PUSH-PULL Method

Figure 2.10 shows that push-pull algorithm over-performs both push and pull algorithms, and spreads the information through all nodes in the fastest way. It approximately takes 15 rounds to inform 4039 nodes.

Here, we recall the study of Karp et al. [11], mentioned in Section 2.2. They proved that push-pull scheme is completed in time $S_n = \log_3 n + O(\ln \ln n)$. In the implementation, they keep a time counter, which represents the age of the information, and it is incremented in each round. Hence, spreading process continues until this counter reaches to the proposed S_n value. When we consider our case, this S_n value becomes approximately 10, and our test results show that 3395 nodes become informed after 10 rounds, so this is actually a vast majority of the network. Then, the remaining 644 nodes become informed in 5 more rounds, because of the randomized scheme.

2.6. Conclusion

This chapter demonstrated how epidemics and information spreading process resembles each other. Information spreading methods, related mathematical models, and topology effect on the spreading process were analyzed. Furthermore, the existing state-of-art methods were implemented to observe information spreading on SNs, especially on Facebook. Results showed that complex epidemic model can be adapted to social networks, and push-pull method is more effective than the other approaches to implement and observe the actual spreading of information on SNs. Ultimately, it was noted that the topology of networks directly affects the information spreading.

There are many other implementation and research domains exist in the context of information spreading. Most of them requires efficient data aggregation algorithms and modelling for dynamic SNs [27, 28]. However, the analysis phase of this theses study does not consider the various reserach domains of information spreading. Instead, the background information provided in this chapter is used to simulate a post dissemination on Facebook in the following chapters.

The rest of this study includes the analysis of Facebook Privacy based on the users' post dissemination process. First, it analyzes the history of Facebook Privacy, and the basics of the Facebook in terms of users' profiles, privacy preferences, and interactions among them, etc. Then, the two privacy leakages of Facebook, which were detected during the mentioned analysis step, are proposed. Finally, the Facebook Privacy Tool, which is proposed for the solution of detected privacy risks is demonstrated, combining with the implementation of information spreading process to visualize interactions among Facebook users.

CHAPTER 3

PRIVACY ISSUES ON SOCIAL NETWORKS: A CASE STUDY ON FACEBOOK

3.1. Introduction

In recent years, SNs such as Facebook, Instagram, Twitter, LinkedIn etc. have become increasingly popular among people of all ages. Especially, Facebook is the most popularly used SN, having approximately 1.97 billion monthly active users, according to the outcome of Statista [1], in April 2017. When we consider the popularity levels of SNs, Facebook is followed by WhatsApp, and Facebook Messenger, respectively. Figure 3.1 demonstrates the number of active users in each month of 2016, and 2017. Results denote that SNs are used by a significant number of people currently, and this number will continue to increase.

	2016	2017
FACEBOOK	1.6	1.97
WHATSAPP	1	1.2
FACEBOOK MESSENGER	0.9	1

Figure 3.1. Monthly Active Users of Most Popular Social Networks (in billion)

Most of the SN users regard their SN accounts as a part of their own lives. Like a common environment, they communicate with their friends, and other people that they even do not know personally. Moreover, they try to keep track of other people's lives at any moment through the instrument of SNs. Hence, many people shape their daily lives to keep up with social life in SNs. Although SNs provide people to be in touch with others easily, they may cause many privacy issues for SN users, if they are unconscious about the possible privacy risks.

SNs have a user-friendly interface, so anyone can create an account, and join the community easily. After getting an account, a user can connect with his/her friends (or

with someone, they do not know), and share any post (information) with them. However, if they do not control the audience of their posts by setting the privacy of them properly, their posts can be visible beyond their expectation. Hence, privacy risks come out for SN users. Based on this problem, we analyzed the privacy risks on SNs (especially on Facebook), and proposed a solution/tool to control SN users' privacy. As a case study, we preferred to work on Facebook, because it's one of the most widely used SN platform today, as mentioned above.

Researchers proposed many solutions, and developed some applications to protect SN users' privacy on Facebook. Some of them were even adapted to the platform by Facebook. However, there are still privacy issues on Facebook, because it is a dynamic platform (it grows continuously with each new friend connection, post/comment/share, etc.). Furthermore, Facebook continuously proposes new updates, especially to solve privacy problems. However, new privacy leakages may arise anytime, because Facebook users may change privacy settings of their profiles/posts, etc., whenever they want.

To cope with this situation, we propose a Facebook Privacy Tool, which will track both the Facebook users' actions, and whole platform to detect any privacy problem, then inform users. We use the information spreading basics to simulate post dissemination on Facebook. Furthermore, as a future work, we propose a new approach for information spreading models to represent human behavior, and real interactions among Facebook users.

This chapter is organized as follows. Section 3.2 includes the history of Facebook Privacy, to explain the domain of this study in detail. Related works on the privacy issues on SNs are given in Section 3.3. Section 3.4 demonstrates the detected privacy problems on Facebook. Proposed solution to the detected privacy problems is presented in Section 3.5. Finally, this chapter is concluded in Section 3.6.

3.2. History of Facebook Privacy

Facebook was launched in 2004, as a student network for Harvard University by Mark Zuckerberg, and his friends. Until 2005, only students from Harvard University could join Facebook. Then, it started to enhance the availability through some high schools, and companies [29]. In 2006, Facebook became publicly available, but the privacy concerns did not change; it was still network-based. Hence, personal data was

accessible through all network by default (the data was public).

Facebook served its users a privacy concerning message for the first time, in 2009. They asked them to select a privacy setting for their personal data, and they did not allow them to access the application without setting their privacy preferences. Boyd et al. [29] discusses about this situation, and shows the options for related privacy settings, as depicted in Figure 3.2. As shown in the figure, Facebook prompts "Everyone" option as default, so many users accepted those default settings without being aware of the privacy risks. Therefore, many data became public, and reachable even by search engines.

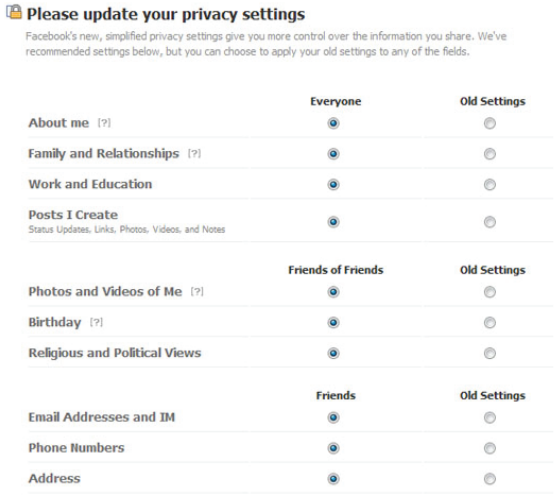

Figure 3.2. First privacy setting page of Facebook

While the mentioned case sparked many debate about the privacy issues on Facebook, a statement coming from Mark Zuckerberg on 8th of January 2010 increased the confusion. He stated that the social norm has been changing in time, since people feel more comfortable with sharing more, and various information with more people openly [29].

As a result of many debates, and critics about possible privacy concerns, Facebook unveiled a new privacy setting page, which is more handy for Facebook users, in May 2010. There was different levels of privacy setting options on the page, such as "Everyone", "Friends of Friends", "Friends Only", and "Other" for each data category [29]. However, it could not prevent privacy leakages; many users' data become publicly available without their awareness. In 2011, Facebook changed some more privacy settings, and those changes allowed people to reach some users' personal data, and profile without being friends [30]. Hence, the privacy of users' almost all data gradually became public

by default settings [31].

Facebook applied a post-based privacy option, which allows users to set specific privacy preferences for each post [32]. Hence, the old network-based privacy structure was removed. In this scope, Facebook proposed four options for privacy settings: (i) Friends : only post owner (PO)'s friends can see it, (ii) Friends of Friends: PO's friends, and friends of his/her friends can see it, (iii) Everyone: anyone can see it, even who is not a Facebook member, and (iv) Custom: PO can create a custom setting by selecting some specific friends or lists, and excluding some others. Although Facebook improve privacy control options, it was claimed that many user information is accessible by third-parties. People were complaining both about privacy problems, and the confusing structure of privacy settings.

Shore et al. [33] worked on every privacy policy of Facebook from 2005 to 2015, and ranked them. Their results show that Facebook Privacy Policies became increasingly incomprehensible, and confusing to understand. They stated that their findings approve the decreasing transparency, and clarity of Facebook's privacy policy. Additionally, they stated that Facebook users have less control over their personal data against third-party access, because Facebook privacy policy contains fewer options to control them [33]. This study confirms that users need some tools, or applications that offers a user-friendly interface, and easy setting/tracking of their privacy.

3.3. Related Works on Privacy Problems of Social Networks

Ho et al. [34] highlights three privacy problems in SNs: (i) lack of user awareness, (ii) lack of flexibility in current privacy tools, and (iii) users do not have a control on what others see about them. To confirm that these problems exist indeed, they conducted an online survey on 200 SN users. This survey consists of twenty-eight questions, including demographic questions (age, gender, etc.), SNs usage questions (friends, profile information, purpose of using SNs, etc.), and privacy concern questions (unauthorized data and intellectual copyrights). Results of this survey confirmed that each of the proposed three problems can be inferred via actual SN users. Hence, they propose a Privacy Framework for SNs, which allows SN users to be aware of the possible privacy risks, understand their own privacy level, and configure this level according to their expectations. This framework classifies SN users with five types: alpha socialisers (who use SNs to flirt, meet new

people, and have fun), attention seekers (who aims to get attention and comments from other people), followers (who use SNs to follow their peers), faithfuls (who use SNs to rekindle old friendships), and functionals (who use SNs for a specific purpose, such as organizing parties, doing charity works, etc.). Then, they evaluate each class of SN users according to their data (i.e. identity, demographic, activity, social network, and profile content), privacy concerns (tendency to share healthy, harmless, harmful, or poisonous data), and profile viewers (best friends, normal friends, casual friends, or profile visitors). As a result of this evaluation, the proposed privacy framework assigns a privacy level to SN users, such as no privacy, soft privacy, hard privacy, or full privacy. Hence, it makes SN users aware of their privacy level, and gives a chance to tune their privacy levels considering the evaluated criteria.

Tuunainen et al. [35] performed a study on privacy concerns and risks of social networking sites, and they especially focused on the factor of privacy awareness. As a case study, they conducted a survey of 210 users of Facebook to observe Facebook users' privacy behavior from two aspects: (i) privacy protection and (ii) information disclosing. The survey consisted of five parts, respectively: (i) users' (respondents') background information, (ii) users' personal information and friends on Facebook, (iii) users' privacy settings, (iv) users' privacy and security concerns, and (v) users' awareness of Facebook Privacy Policy. Results strongly showed that most of the users do disclose a significant amount of sensitive information of themselves. In addition, they are not sufficiently aware of the visibility of their information to people, who they do not actually know. Furthermore, many users do not know or understand the Facebook Privacy Policy, and the terms of use of Facebook. Unfortunately, many of the users even noted they were not aware that Facebook can share their sensitive information with third parties. Rewardingly, they remarked as a feedback to this survey that they will pay more attention to protect their privacy while using social networking sites hereupon.

Talukder et al. [36] claims that SN users may protect their sensitive information by tuning the privacy settings of them properly, but that information may still be exploited by adversaries, via prediction techniques. That is why, properly tuning the privacy settings of sensitive information is not enough to protect its privacy, so SN users need more advanced solutions.

Talukder et al. [36] proposed "Privometer", which is a tool, used for measuring the level of privacy leakage for a specific user, and directs him/her to self-sanitization process by recommending some options to lessen this privacy leakage in his/her profile. In the

scope of this tool, they assume that a potentially malicious application has been installed by some friends of a Facebook user, so it may access those friends' sensitive information. However, it cannot directly access to the user's sensitive information, because he/she did not install that malicious application. Even so, the malicious application may still access his/her sensitive information by using an inference algorithm. Considering this situation, Privometer checks the best-known models [37], that can be used for this inference, and tries to predict which algorithm is used by the malicious application. To do so, Privometer selects the model, that infers user's sensitive information most accurately. After that, it measures the level of user's privacy leakage, and demonstrates the result by a friends list, ranked by the detected threat from each of them. Additionally, it proposes some self-sanitization choices (recommendation) for the user to decrease the level of his/her privacy leakage.

Wilson et al. [38, 39] focus on the analysis of real interactions among SN users. They propose the idea that commonly referred social links does not actually represent real interactions. If so, SN users who have a social link with a post owner, had to interact with that post because of the corresponding link between them. However, this is not the case on SNs, and this study proves that few linked users create an interaction graph with a big diameter, and small number of super nodes [38, 39]. They used a real dataset, created by them, which consists of real Facebook user traces, and so includes real interactions from Facebook users. Hence, they created an interaction graph based on this dataset. This graph includes same nodes as social link graph, but it only takes a subset of the links. Some of their findings are listed below [38, 39]:

- "Most users have no interaction with up to 50% of their friends."
- "For the vast majority of users ($\sim$ 90%), 20% of their friends account for 70% of all interactions."
- "Nearly all users can attribute all of their interactions to only 60% of their friends."

Wilson et al. [38, 39] gives a reasonable background for this thesis study, because they provide percentage values from real data, which can be used in the experimental work.

Analysis of interaction among SN users has begun to be a popular topic for researchers in recent years. In this scope, James et al. [40] proposes the idea of "dual privacy decision" for SN users' behavior. This idea provide SN users choose what information to release (information), and who can view it (interaction). This study includes

four key motivations as following [40]: (i) seeking information, (ii) socializing, (iii) expressing oneself to others, and (iv) meeting social expectations or pleasing others. Hence, they try to handle the privacy concerns on SNs from two aspects: (i) information, and (ii) interaction, and provide an analysis of privacy management schemes for SNs. These schemes are classified as "information managers" and "interaction managers". Information managers have many friends, while interaction managers have less. James et al. [40] performs some tests on these schemes, and show that socialization, self-expression and pleasing others influence information and interaction behaviors on Facebook, while information seeking does not influence. Hence, this study confirms that apart from social links, there are also other factors that affects interactions among SN users.

Dong et al. [41, 42] propose a privacy decision-making tool, which considers the SN users' behavioral model. By using a prediction model, this tool gives personal advices for SN users according to their anticipated preferences to help them in their privacy decision-making process. The prediction model in this study based on a set of psychological and contextual factors [41, 42], such as the trustworthiness of the requester/audience, the sharing tendency of the user, the sensitivity of the information, the appropriateness of the request/disclosure, and several traditional contextual factors. To learn the influence of all these factors, they use a binary classification model. Moreover, they rank them according to their chi-squared statistics, and information gain. By using the model, proposed in this study, people can handle the tradeoff between the potential benefit, and risk of information disclosure decisions on SNs.

Aghasian et al. [43] propose a framework to measure the online social network users' privacy disclosure scores (PDS), considering multiple SNs. They defined the main factors that affect users' privacy as sensitivity and visibility. Based on these two factors, they proposed a scoring function, which also considers a set of common personal attributes, either with the form of structured (i.e. username, age, etc.) or unstructured (i.e. messages, images, etc.) data. After obtaining the framework attributes, they first compute the sensitivity of each user, using the determined values by Srivastava et al. [44]. Then, they compute the visibility of each user, considering the three factors, respectively: (i) ease of accessibility, (ii) difficulty of data extraction, and (iii) data reliability. After calculating the effect of each three factor, they use a set of fuzzy rules to find the overall visibility score for the attributes of each user. Specifically, they use a fuzzy inference system, based on the Mamdani fuzzy inference [45]. Hence, they measure the sensitivity and visibility of each user, and then they combine these results to calculate users' privacy

disclosure scores. They stated that, the more this score is, the more likely a user in a privacy risk and disclosure. Their experiment results showed that users' privacy disclosure scores highly depend on the amount of information, disclosed by the users themselves.

3.4. Currently Detected Privacy Problems on Facebook

During our research period for this study, we examined Facebook Privacy Basics [46], which are announced officially by Facebook, and the whole mechanism behind the post dissemination process on Facebook. After performing real tests on a small group of official Facebook users' profiles (a group of 5 people), we detected two crucial privacy leakages, which are explained below in detail:

1. **Comment Owner (CO)'s interaction may be visible to a bigger area than expected:** Suppose that a Facebook user interacts with a post (so becomes a CO), by leaving it a comment or like, and the privacy setting of this post is "Friends only". After a while, the PO changes this setting, and enhance its audience beyond friends. In this case, Facebook does not inform the corresponding COs, who have already interacted with the post. Hence, interactions on this post become visible to more people than the COs' expectation. This privacy problem is also mentioned in the officially announced Facebook Privacy Basics [46], and Facebook does not provide any solution for it.

2. **COs' interactions may be inaccessible by themselves:** Suppose that Facebook user A, and B are friends, and user C is B's friend. User A creates a post by setting its privacy to "Friends only", then tags user B on this post. Hence, post privacy of the created post becomes "Friends and B's friends". Then, user C interacts with this post. After a while, user A removes user B's tag (untagging operation on Facebook). In this case, user C is not able to reach his/her comment or like anymore, because the privacy setting of the post became "Friends only" again, and it can be seen only by A's friends. In this case, user C even cannot see any record of his/her interaction on his activity log page. This situation creates a privacy problem for user C, because he/she cannot see his/her own comment or like on the post, while some other people can still see it. This privacy problem on Facebook is not covered in Facebook Privacy Basics [46], and detected during our analysis, and real tests.

Hence, it is important to present this problem to Facebook users, and make them aware of the possible risks.

3.5. Solution Proposal to Detected Privacy Problems on Facebook

A novel Facebook Privacy Tool is proposed in this section. This tool is expected to dynamically track users' action logs, check the platform for any change in users' privacy settings, detect privacy issues, continuously inform users, suggest some solutions to detected problems, and direct users to protect their privacy. Additionally, it will also create awareness about possible privacy risks, that Facebook users may face with. Main facilities of the proposed tool are listed below.

Fa. PO is informed about the spreading area of a specific post, based on its privacy setting.

Fb. A possible interaction graph of a specific post, that includes users who may interact with this post, is demonstrated to PO.

Fc. When a PO changes the privacy setting of his/her previously created post, COs who have already interacted with this post are informed about the change of privacy setting.

Fd. COs can set some rules via this tool to protect their interactions (comments/likes) on other users' posts (i.e., "Automatically delete my interaction from a specific post, if its privacy settings is changed by its owner.").

3.6. Conclusion

This chapter presented the place and importance of SNs today, and highlighted that SNs may create some privacy risks for their users. To remark the case study of this thesis, we provided the history of Facebook Privacy, and then explained the current privacy issues on Facebook, which were detected during our research period. Finally, we proposed a solution to those problems, which can be implemented as a real Facebook tool, and serve as a privacy-checking assistant to Facebook users.

In the remaining part of this theses study, we did not develop this proposed tool as a real application, since its applicability depends on the Facebook itself, as it will be

explained in Section 4.5. However, we implemented the required tests on SNAP dataset, and created a simulation to represent the real tool. Therefore, the following chapter will first propose our experiments, and then show a possible design, and implementation of the proposed tool.

CHAPTER 4

VERIFICATION OF THE PROPOSED SOLUTION FOR PRIVACY PROBLEMS ON FACEBOOK

4.1. Introduction

This section includes the experimental work of this theses study, and focus on the implementation, verification, and simulation of the proposed tool. To satisfy the facilities of the tool, mentioned in Section 3.5, we focused on three research questions (RQ), presented below respectively.

RQ1. Does POs' popularity level affect the spreading area of their posts, and the number of interactions they may get from other users?

This research question is related to Fa. and Fb., stated in Section 3.5. We consider the popularity level of a PO as the number of his/her friends (i.e. the number of directly connected neighbor on SNAP). To analyze the relation between popularity level of a PO, and the number of interactions that his/her post may get, we classified the POs according to their popularities. Hence, we took this classification into account while performing our experiments.

RQ2. Does any change in privacy setting of a post affect COs' decision on keeping their interaction on this post?

This question focuses on the facility Fc., and caused by POs' privacy preferences. POs can select one of the following three choice in our experimental setup: (i) Friends: Only PO's friends can see the post, (ii) Friends of Friends: PO's friends, and friends of the tagged friends can see the post. This choice becomes active when PO performs at least one tagging. Tagging someone on Facebook means that the PO selects some of his/her friends, and put a tag for him/her to add his/her friends as audience of the post. Tagging is performed with a real-time decision by POs. However, to perform the experiments efficiently with the static data, we assumed that a POs has a small probability to tag any of his/her friends for their posts. To deal with a reasonable run time in our tests, and define

a causal value, we decided to select this probability as 20%, based on the study of Wilson et al. [38, 39] (iii) Public: Anyone can see the post, even someone who is not a Facebook member. After the PO select one of these choice as the privacy setting of his/her post, he/she may change this setting whenever he/she wants. In this case, COs may want to control their interaction on the related post.

RQ3. Do the rules defined by COs for a specific post affect the number of interactions that the post may get?

This research question considers the facility Fd. If COs define some rules while leaving a comment/like on a post, these rules may help them control their interactions (so control also their privacy) in a better way. If this is the case, COs may remove or change their interaction on a specific post, and the number of interactions on a post may vary in time.

Considering all of these, we tried to develop the proposed tool, and simulate it according to experiment results. An analysis of the three research questions, and experiments on SNAP dataset are presented in Section 4.2. Section 4.3 demonstrates a simulation of the proposed solution. A possible implementation of the proposed tool is given in Section 4.4. Section 4.5 discusses the applicability of the proposed tool. Finally, this chapter is concluded in Section 4.6.

4.2. Experimental Work

This experimental work covers the analysis of whole dataset to search for the answers of mentioned three research questions. For this purpose, we first tried to discover the popularity levels of each node in the dataset. We accepted the popularity of a node, as the number of their neighbors (so degree of the node represents its popularity level). Hence, we found the degree of each node, and then sort them according to their popularity levels in descending order, in a list (L_{pop}). Then, we created ten different popularity class, and classified the whole nodes in the dataset according to their popularities, as following: (i) 1^{st} class: first 10% of nodes in L_{pop} (so this class keeps the most popular 10% nodes in the dataset), (ii) 2^{nd} class: after first 10% popular nodes, take the next first 10% remaining in L_{pop}, etc. Hence, 10^{th} class represents the least popular nodes in dataset. As a result, we created ten different groups of nodes (Gr_{pop}) according to their popularity, and each group $g \in Gr_{pop}$ contains approximately 400 nodes, accordingly.

For each group g, we performed the steps given below:

1. Create a sample (S_{pop}) from g, that contains 100 randomly selected nodes.
2. For each node Nd in S_{pop}:

   2.a. Find the Friendship graph (FG) of Nd, and record this graph.

   2.b. Find a possible interaction graph (IG) for Friends case, and record it.

   2.c. Determine the friends to tag for the Friends of Friends (FoF) case as mentioned above.

   2.d. Create FoF graph of Nd and record it.

   2.e. Find a possible IG for FoF case and record it.

Step 2.a. and 2.b. was required to answer our RQ1. We would like to observe the change in the size of friendship graph and the number of interactions that the related POs get according to their popularity levels.

In Step 2.a., POs' friendship graphs were found based on the relations in our dataset. In this graph, friends were represented as nodes, and friendship relations were represented as edges.

In step 2.b., size of the possible IG for each Nd was measured. Based on the study of Wilson et al. [38, 39], the probability of a node to interact with his/her friends' post was taken as 0.2. Hence, the possible IG was created by assigning a decision value (i.e., decision for interacting with the post) for each node in Friends graph of Nd. Decision value was calculated by assigning a random value (1 or 0) with probability of 0.2 (p_F) for yes (1), and 0.8 for no (0). This step was repeated 50 times for each node in the FG of Nd, and the average number of interactions was calculated. This average value was used to specify the size of the possible interaction graph. Interaction graph includes the nodes that have decision value 1, and the edges come from the friendship relation among these nodes. The number of nodes that the graph include was recorded as the size of this graph.

Step 2.c. was a preparation process to create FoF graph of the related nodes, since we should first specify some friends of the node to tag. Then, based on the determined friends, we created the FoF graph of the nodes by extending the FG with them.

In step 2.e., first of all, friends who will remove their interaction when the privacy setting of the post changes from Friends to FoF was stochastically calculated. Then, the change in the spreading area of the post was observed, while testing for different

probability values for a node to remove his/her interaction. A removal decision value was assigned to each node, based on the probability of withdrawal (p_w). The whole process was repeated for p_w values of 10%, 20%, 30%, 40%, and 50%, respectively. In each trial, IG of Friends was updated, and extended with the addition of new nodes from FoF, after finding the removal decisions. For the extension operation, a decision value was assigned to interact with each FoF node (same with p_F, so $p_{FoF} = 0.2$), and the ones with decision 1 were added to IG. Final graph represented the interaction graph of FoF. Hence, the size of this graph was recorded as the number of interactions (nodes) for FoF case.

Actually, we implemented Step 2.c., 2.d., and 2.e. to answer our RQ2 and RQ3. We tried to analyze the effect of a privacy change in an existing post to COs' behaviors on this post. In addition, We simulated the privacy awareness of a CO with p_w values, and thus observed the change of COs' behavior in different popularity levels.

Table 4.1 shows the change in number of interactions for each popularity class for different p_w values. Rows in this table start with the most popular nodes (1^{st} class), and continue through the least popular (10^{th} class). For each row, corresponding columns represent the followings:

- $len(FG)$: Expected average number of nodes in FG of the corresponding popularity class (from Step 2.a).
- $I(F)$: Expected average number of interactions for Friends case of the corresponding popularity class (from Step 2.b).
- $len(FoFG)$: Expected average number of nodes in FoF case of the corresponding popularity class (from Step 2.d).
- $I(FoF)$: Expected average number of interactions for FoF case (from Step 2.e). This column was split into six sub-columns; one for the number of interactions without any removal ($p_w = 0$), and others with removals to observe the change in different p_w values.

Results show the following inferences:

- Considering RQ1, we can say that popularity level of a PO directly affects the size of his/her FG and $FoFG$, and the number of interactions that any of his/her post may get.

Table 4.1.: Change in number of interactions vs p_w

L_{pop}	$len(FG)$	$I(F)$	$len(FoFG)$	$I(FoF)$					
				$p_w = 0$	$p_w = 10\%$	$p_w = 20\%$	$p_w = 30\%$	$p_w = 40\%$	$p_w = 50\%$
1^{st}	163, 31	32, 67	475, 66	121, 23	85, 10	88, 65	85, 37	80, 49	78, 96
2^{nd}	87, 93	17, 39	354, 91	84, 66	65, 45	67, 44	65, 36	62, 95	62, 15
3^{rd}	58, 56	11, 68	319, 29	72, 94	60, 17	61, 18	60, 18	58, 85	57, 91
4^{th}	41, 92	8, 39	238, 94	54, 21	45, 14	45, 95	45, 33	43, 95	43, 75
5^{th}	30, 55	6, 10	268, 36	58, 30	51, 78	52, 47	51, 74	50, 84	50, 67
6^{th}	22, 57	4, 60	182, 04	39, 78	35, 10	35, 56	35, 19	34, 56	34, 20
7^{th}	16, 91	3, 31	160, 34	34, 24	31, 00	31, 25	31, 12	30, 58	30, 39
8^{th}	14, 43	2, 52	179, 13	37, 40	35, 07	35, 37	35, 04	34, 73	34, 62
9^{th}	7, 83	1, 62	105, 46	22, 05	20, 62	20, 74	20, 74	20, 41	20, 27
10^{th}	3, 94	0, 78	101, 34	20, 54	20, 02	20, 10	20, 04	19, 86	19, 91

- Number of friends who can see the post increases, when we expand privacy setting of the post from "Friends" to "FoF", as can be seen from the columns $len(FG)$ and $len(FoFG)$.

- Effect of the proposed tool can be observed when we consider the case of removals ($p_w = \{10\%, 20\%, 30\%, 40\%, 50\%\}$). Spreading area of the post becomes bigger, when we change the privacy from "Friends" to "FoF", so the number of interactions it may get also increases dramatically. However, if COs define some rules for their interactions, increase rate in number of interactions may be smaller than PO's anticipated number. This property helps people take a precaution to protect their privacy by setting rules for their interactions on others' posts. For instance, 3^{rd} row in Table 4.1 shows that although PO's expected number of interactions for FoF case is approximately 73, this number decreases to approximately 58 if users set some rules to protect their privacy with probability of 50% ($p_w = 50\%$). This result gives an acceptable result to our RQ2 and RQ3. The impact of privacy changes in the existing posts on Facebook users' behaviors significantly increases, while their privacy awareness is increasing. Hence, they may start to protect their interactions on other users' posts by deleting them in case of any audience expansion.

To observe the inferred results in a better way, we also performed the same experiment on the whole nodes in dataset, instead of taking a sample of the graph. Table 4.2 shows the experiment result in the case of using whole nodes in dataset. As seen from the table, there is no significant difference but the $I(FoF)$ values shows a more smooth decrease against increasing p_w values.

All in all, we can say that this experiment has an impact on creating awareness for

Table 4.2.: Change in number of interactions vs p_w (case of using whole nodes)

L_{pop}	$len(FG)$	$I(F)$	$len(FoFG)$	$I(FoF)$					
				$p_w = 0$	$p_w = 10\%$	$p_w = 20\%$	$p_w = 30\%$	$p_w = 40\%$	$p_w = 50\%$
1^{st}	164, 21	32, 94	458, 78	117, 85	88, 60	85, 32	81, 85	77, 25	75, 47
2^{nd}	88, 88	17, 85	391, 88	92, 50	76, 67	74, 79	73, 18	70, 60	69, 54
3^{rd}	58, 92	11, 79	332, 85	75, 72	65, 42	64, 22	63, 17	61, 26	60, 66
4^{th}	41, 50	8, 34	288, 87	64, 07	57, 02	56, 13	55, 31	54, 15	53, 58
5^{th}	30, 59	6, 11	262, 64	57, 02	51, 85	51, 33	50, 73	49, 77	49, 53
6^{th}	22, 66	4, 54	224, 45	48, 10	44, 41	43, 99	43, 51	42, 85	42, 63
7^{th}	16, 91	3, 39	182, 26	38, 79	36, 13	35, 81	35, 48	34, 91	34, 78
8^{th}	12, 23	2, 46	173, 96	36, 35	34, 53	34, 31	34, 12	33, 78	33, 50
9^{th}	8, 05	1, 61	134, 34	27, 76	26, 66	26, 57	26, 38	26, 12	26, 06
10^{th}	3, 91	0, 78	106, 97	21, 68	21, 33	21, 21	21, 20	21, 04	20, 98

Facebook users by highlighting that the privacy setting of a post and the PO's popularity level are so important to estimate its spreading area. Keeping this in mind, Facebook users can control their interactions on others' posts by tracking their privacy settings, and so predicting to whom (or to how many people) their interactions may be visible.

4.3. Simulation of the Proposed Tool

This section presents a basic simulation, which provide a visual representation of the proposed tool. This simulation tool works with eleven steps on SNAP dataset. Details of each step is explained below, including the details of a complete run of the tool. The example run works with an average-popular node as a PO, and the probability values for p_F, p_{FoF}, and p_w are accepted as 0.2.

Step 1 - Visualization of dataset: First step covers the visualization of the whole network (dataset). Figure 4.1 shows this visualization of SNAP Facebook Dataset, which includes 4039 nodes. This figure also demonstrates the Graphical User Interface (GUI) of the simulation tool. User clicks the buttons at the bottom of screen, and continue his/her actions with the following steps.

Step 2 - Selecting a PO: To observe a complete run of the simulation tool, an average-popular node is selected as the PO (with ID 766 from the 5^{th} popularity class).

Step 3 - Creating a post (Shared with "Friends" as default): A default post is created.

Step 4 - Spreading the post & visualizing the spreading area (Friends case): Post is spread through PO's friends according to push-pull method, and then the spreading

area is visualized as in Figure 4.2. Resultant size of the friendship graph is 36 ($len(FG) = 36$).

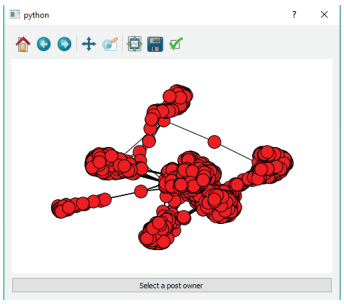

Figure 4.1. SNAP Facebook Graph (Shows all nodes in dataset)

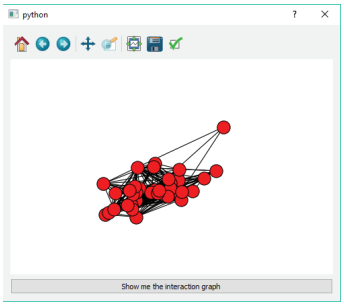

Figure 4.2. Spreading area of the post (Privacy Setting: "Friends")

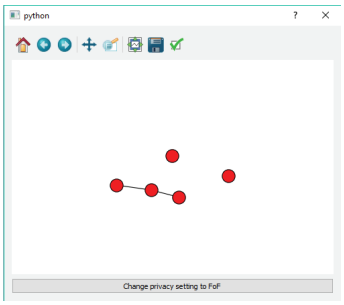

Figure 4.3. Interaction Graph of the post (Privacy Setting: "Friends")

Step 5 - Creating and visualizing a possible interaction graph (Friends case): A possible interaction graph is created for the post, based on the method explained in Section 4.2. Then, created graph is visualized as shown in Figure 4.3. Number of interactions for "Friends" case is 5 ($I(F) = 5$).

Step 6 - Changing privacy setting of the post from Friends to FoF: Privacy setting of the post is changed to FoF. That is why, PO expects more number of interactions on his/her post than the previous case.

Step 7 - Spreading the post & visualizing the spreading area (FoF case): Post is spread again by using push-pull method, but this time it reaches to PO's friends, and friends of the tagged ones with $p_{FoF} = 0.2$. Then, the updated spreading area is visualized, as in Figure 4.4. Unsurprisingly, number of nodes that can see this post increased as shown from the difference between Figure 4.2, and Figure 4.4. Size of the spreading area for FoF case is 48 ($len(FoF) = 48$).

Step 8 - Creating and visualizing a possible interaction graph (FoF case): A new interaction graph is created for FoF case, and visualized as shown in Figure 4.5. It can be seen from the difference between Figure 4.3, and Figure 4.5 that there is a noticeable decrease in the number of interactions, when we change the privacy setting from "Friends" to "FoF", although the spreading area for FoF case is bigger than the one in Friends case, as mentioned in Step 7. This situation demonstrates that there is a possibility of getting less number of interactions when we enhance the audience of a post (i.e. $I(F) > I(FoF)$), even if the spreading area of the post is increased (remember that

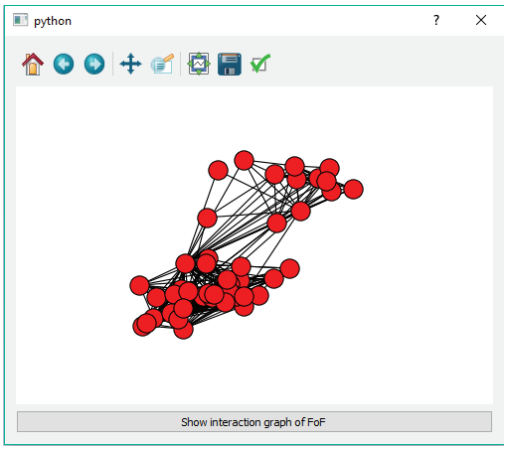

Figure 4.4. Spreading area of the post (Privacy Setting: "FoF")

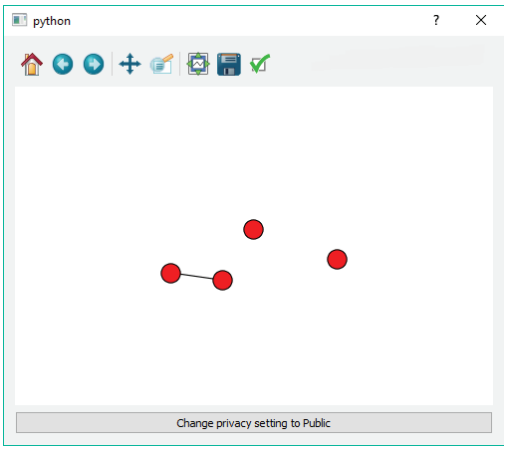

Figure 4.5. Interaction Graph of the post (Privacy Setting: "FoF")

the privacy awareness of COs is 0.2 in the experiment, so $p_w = 0.2$). In this case, PO's expectation of getting more interactions on his/her post cannot be satisfied. This is not the case for all simulation scenarios, but it should be remarked that there is a possibility to meet with a decrease in the number of interactions, if the COs' privacy awareness is increased. As a result, number of interactions decrease to 4 in this example ($I(FoF) = 4$).

Step 9 - Changing privacy setting of the post to Public: Privacy setting of the post is changed to "Public".

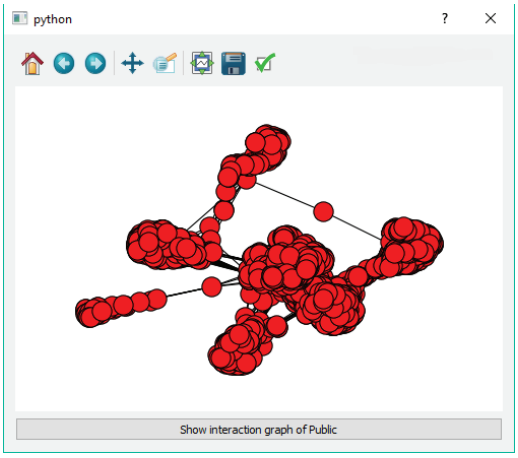

Figure 4.6. Spreading area of the post (Privacy Setting: "Public")

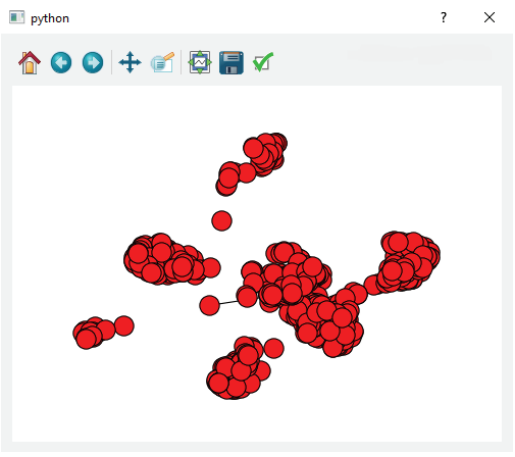

Figure 4.7. Interaction Graph of the post (Privacy Setting: "Public")

Step 10 - Spreading the post & visualizing the spreading area (Public case): Post is spread through whole network by using push-pull method. Updated spreading area

of the post for “Public” case is visualized as shown in Figure 4.6. In this case, size of the spreading area becomes equal to the size of network, so it is 4039.

Step 11 - Creating and visualizing a possible interaction graph (Public case): A possible interaction graph for the “Public” case is created and visualized as in Figure 4.7. Number of interactions for this case becomes 803.

4.4. Possible Implementation of the Proposed Tool

This section includes a mockup design of the proposed tool, to present how the tool will work. This mockup was created by using the Proto.io Website [47], and the link [48] includes a video which shows a complete run of the mockup. The mockup has eight main screens (please see Appendix A), and the screens represent only dummy values for the size of spreading area, or the number of interactions, so it does not actually work with the real data. Work flow of the mockup is presented in Figure 4.8, and the detailed explanation for each main screen is provided below.

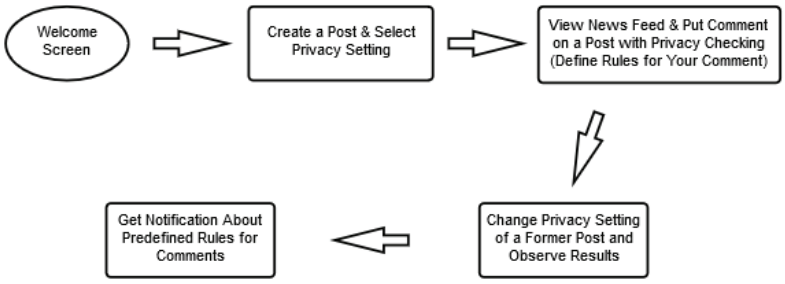

Figure 4.8. Flowchart of the Mockup

1. **Welcome (Figure A.1):** First main screen welcomes users, and inform them about the scenario embedded in the mockup, addressed privacy issue, and the content of images. After that, directs users to follow the flow of mockup, respectively.

2. **Create a Post (Figure A.2):** Post creation is simulated by using a dummy Facebook screen. In this screen, post is ready as default, and the PO is expected to proceed with this default post.

3. **Select a Privacy Setting (Figure A.3):** This screen directs the PO to select the privacy setting of his/her post as "Friends". According to this setting, spreading area of this post is visualized via a sub screen, as shown in Figure 4.9. It can be seen from the figure that dummy numbers for the size of spreading area for each case are given at the top, with red, orange, and green colors. Furthermore, corresponding area are represented with circles. In the real application form of this tool, those number will show the real data for each PO.

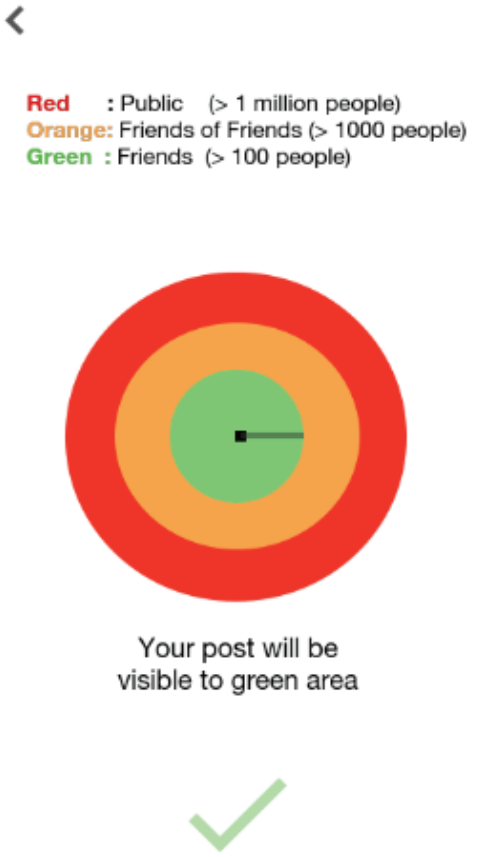

Figure 4.9. Mockup Screens for "Select Privacy Setting" (Friends)

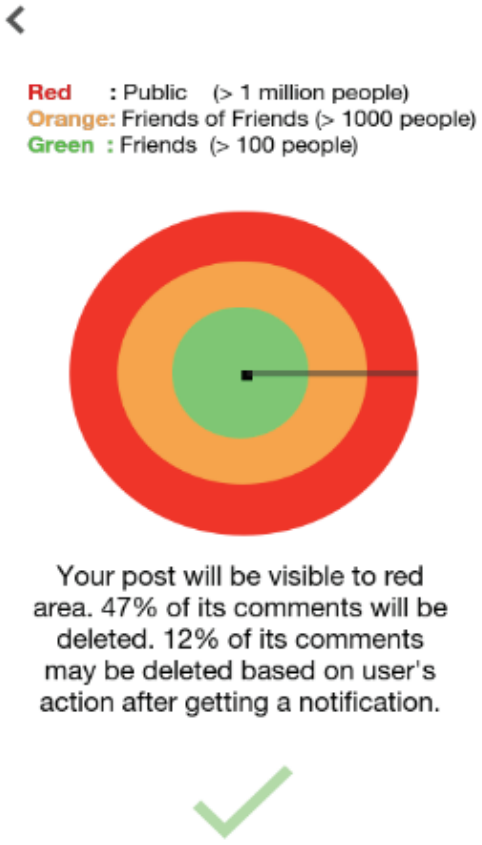

Figure 4.10. Mockup Screen for Privacy Change to Public

4. **View News Feed (Figure A.4):** This screen simulates some dummy posts which were created by PO's, or CO's friends. This property is referred as "News Feed" on Facebook.

5. **Put a Comment (Figure A.5):** This screen directs the user to interact with a specific post, by leaving a comment, and so the user becomes a CO. A dummy com-

ment is already placed in this screen, so the CO just need to proceed to the next screen.

6. **Check Privacy (Figure A.6):** When the CO wants to send his/her comment, he/she is asked to check the privacy of the comment before putting it. Hence, this screen shows a message on the screen which asks whether the CO prefers to set a rule for this comment to protect its privacy. If so, the mockup proposes some rule options for him/her (i.e. "Send me a notification!", "Delete my comment without notifying me!", and "Delete my comment and notify me!"), and he/she can select one of these rules.

7. **Change Privacy Setting of Former Posts (Figure A.7):** This screen represents PO's previously shared posts. PO is directed to select a specific post which is created by him/her before, and then try to change the privacy setting of it to public. Before applying the corresponding change in privacy setting, mockup demonstrates the newly created spreading area for this change, as shown in Figure 4.10. PO can see the result of this change in the number of interactions on this post, as percentage values at the bottom of the screen. These results are affected by the rules created by COs of the post. Numbers in Figure 4.10 are dummy, that were used for only this simulation, and they should reflect the real data in the original application.

8. **Get Notification from Tool (Figure A.8):** This screen demonstrates a notification that the tool sends to COs in case of a change in any of the post which they interacted with. Thanks to this notification, COs can control the privacy level of their interactions, affected by the changes in privacy settings.

4.5. Applicability of the Proposed Tool

Applicability of the proposed tool can be considered in two aspects: (i) a centrally managed approach, and (ii) an application-based approach.

1. **Centrally managed approach:** In this approach, we can consider the proposed tool as a service, served by Facebook officially. To do so, Facebook should allow COs to reach all their personal activity logs, even the ones that is no longer accessible by the CO. Considering the first research question (RQ1), proposed in the beginning of

this chapter, Facebook should send a kind of notification to COs, if a privacy change occurs on a post that they have already interacted. For the case of RQ2 and RQ3, Facebook should give COs access to their all activity logs although the visibility of the corresponding posts has been changed. This approach is handy for Facebook users, but causes an extra work-load, and complexity for Facebook.

2. **Application-based approach:** This approach considers developing the proposed tool as a Facebook App. If this happens, we should expect all Facebook users install this application to their accounts, because we should have access to their all post-based information to make this application produce realistic results. Otherwise, we cannot monitor the whole possible privacy risks. If all Facebook users install this tool as a Facebook application, the tool can inform them about any privacy risk, as presented in the mockup, in Section 4.4.

4.6. Conclusion

This section covered experimental work for the implementation of the proposed tool, provided in Chapter 3. The experiments mainly focused on three research questions. First research question considered the effect of POs' popularity on the spreading area of their posts, and the number of interactions they may get. Second one examined how does any change in the privacy setting of a post affect COs' decision on whether to still interact with the corresponding post or not. Last research question was related to the analysis of the effect of COs' defined rules on the number of interactions that the corresponding post gets. To observe these concepts, experiments was performed on Facebook SNAP Dataset [2]. Results showed that POs' popularity level directly affects the size of the spreading area for their posts, and the number of interactions the related posts may get. Additionally, any change in the privacy setting of a post may affect COs' decision about their interactions on this post, and the number of interactions for a post may be affected by the rules that were defined by COs.

In addition, a simulation which works for any specific node was presented with a complete run of a node, which has an average popularity. In this simulation, it was proved that the number of interactions for a specific post may decrease, when we enhance its audience (although owner of the post expects to get more interaction). This result shows

that COs can control, and protect the privacy of their interactions, if they are conscious enough.

Finally, a possible implementation of the tool as a real-world application was demonstrated as a mockup, to visualize the proposed solution in a better way. Furthermore, the applicability of the tool was discussed to emphasize its feasibility.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1. Conclusion

In this thesis study, an analysis of information spreading and privacy problems on SNs was performed with a case study on Facebook. It has been increasingly difficult to observe this analysis, because SNs have a dynamic structure, and density of the information exists on SNs continuously increases. In these circumstances, it has been getting harder to control the privacy of information on SNs. Hence, the main objective of this study was to examine the spreading process of information on SNs, while controlling the privacy level of SN users' sensitive information, and provide some solution to increase this privacy level by creating awareness about the detected privacy problems.

For this purpose, the first step was analyzing the background of information spreading to comprehend its underlying mechanism, and the state-of-art methods used in the literature. It was noticed that, the topology of network is an important factor on the spreading process of information. Considering the topology of SNs, it was pretty certain that the networks, that this study focus on, have frequent updates, and inherits a dynamic structure. Hence, push-pull method, provided in Section 2.2, was selected to implement in the simulation of post dissemination on Facebook, as proposed in Section 4.2 and 4.3.

The study continued with analyzing the privacy on Facebook. First, the history of Facebook Privacy was examined, and the related works on the privacy issues on SNs was reviewed. Then, Facebook was investigated to observe users' privacy levels. It was detected that there are currently two important privacy leakages on Facebook, which were stated in Section 3.4. After that, a solution to solve the detected privacy problems was proposed.

The proposed solution mainly focused on creating awareness for the Facebook users by providing them to observe the spreading area of their posts, and control their interactions on other users' posts. The control mechanism provided users to set some rules for their specific interactions before posting them. Rules was related to control the

privacy of a specific interaction by tracking the privacy setting of the post on which this interaction appears. If the privacy setting of the post is changed by its owner, users who interacted with this post is informed, or their interactions are automatically deleted from the post, according to the rule they defined.

Experimental work covered the analysis of whole dataset to observe the spreading area of a post, and the number of interactions it gets according to different privacy settings. Results showed that (i) the spreading area of a post is affected by its privacy setting, (ii) the popularity level of the post owner directly affects the number of interactions that any of his/her post get, and (iii) if users become aware of the privacy risks and try to protect the privacy of their posts/interactions using the rules provided by the proposed solution, they can control their posts/interactions, and increase their privacy level.

Finally, a simulation was demonstrated in Section 4.3 and a mockup was presented in Section 4.4 to express the proposed solution visually. A complete run of the proposed solution with a scenario was included in these demonstrations to express it better. Furthermore, applicability of the tool was discussed to highlight that it is a convenient and feasible solution.

5.2. Future Work

Understanding the underlying structure of the information spreading among SN users is a crucial matter to protect privacy on SNs. That is why, analyzing the information spreading model on SNs is an important issue for us. During this thesis study, we realized that information spreading process on SNs is mostly affected by SN users' behavior, and the existing information spreading models are not sufficient to represent the spreading behavior today. To develop a real-world information spreading model, we first analyzed the requirements, such as the popularity of the information source, strength of relations among users, content of the information, personal interests, and privacy preferences, etc. [19]. We believe that all these requirements should be considered as a factor in the development of an information spreading model that will be proposed.

Some researchers [20, 21, 22, 25] have already proposed the modified versions of SIR model or new approaches by using information cascades, as stated in Section 2.4, to make their model more realistic. Bao et al. [20] and Cordasco et al [22] modified the SIR model to adjust its states to current environment (i.e. adding an aware state, or

dividing the Infected state into two which represent the people who belived/not believed in the information). Furthermore, Tong et al. [25] proposed using cascades to model the information spreading.

Considering all the approaches in the literature, we designed a hybrid information spreading model, which combines the states proposed by Bao et al. [20] and Cordasco et al, and focuses on the transitions between these states. Figure 5.1 shows the states and transitions of the proposed information spreading model [19]. Hence, we propose to use five states: (i) Ignorant: all users in a SN are assumed to be ignorant initially, (ii) Aware: a user becomes Aware when he/she gets the information, (iii) Positive Infected: a user becomes Positive Infected and starts to infect others positively if he/she believes in the information, (iv) Negative Infected: a user becomes Negative Infected and starts to infect others negatively if he/she does not believe in the information, and (v) Removed: a user becomes Removed when he/she stops the spreading. Regarding to transitions between states, they will represent user-centric threshold values to define a state change. Further explanation about each transition can be found in the related paper [19].

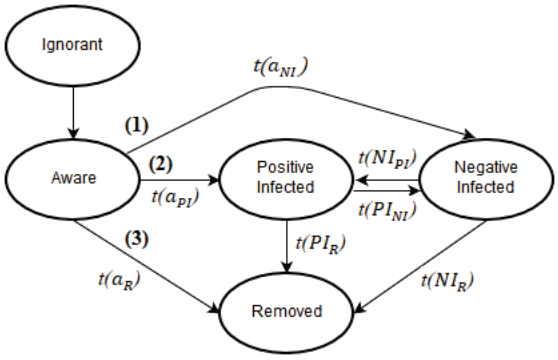

Figure 5.1. A Hybrid Information Spreading Model

All in all, we propose to consider the idea of information cascades, and the requirements mentioned above to include users' behavioral effect in all transitions of the proposed model. Hence, we will measure different threshold values for each user to perform each transition, so that we can represent the behavioral effect of a user in the information spreading model.

REFERENCES

[1] Statista. Number of social media users worldwide from 2010 to 2021 (in billions), 2017. https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

[2] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection, June 2014. http://snap.stanford.edu/data.

[3] Alan Demers, Dan Greene, Carl Hauser, Wes Irish, John Larson, Scott Shenker, Howard Sturgis, Dan Swinehart, and Doug Terry. Epidemic algorithms for replicated database maintenance. In *Proceedings of the Sixth Annual ACM Symposium on Principles of Distributed Computing*, PODC '87, pages 1–12, New York, NY, USA, 1987. ACM.

[4] D Daley and D.G. Kendall. Stochastic rumours. 1, 03 1965.

[5] C. Nowzari, V. M. Preciado, and G. J. Pappas. Analysis and control of epidemics: A survey of spreading processes on complex networks. *IEEE Control Systems*, 36(1):26–46, Feb 2016.

[6] Compartmental models in epidemiology, 2017. https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology.

[7] Troy Tassier. Simple epidemics and sis models. In *The Economics of Epidemiology*, chapter 2. SpringerBriefs in Public Health, 2013.

[8] W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 115(772):700–721, 1927.

[9] Fred Brauer. The kermack-mckendrick epidemic model revisited. *Mathematical biosciences*, 198(2):119—131, December 2005.

[10] Norman Bailey. *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London, 1975.

[11] R. Karp, C. Schindelhauer, S. Shenker, and B. Vocking. Randomized rumor spreading. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 565–574, 2000.

[12] Flavio Chierichetti, Silvio Lattanzi, and Alessandro Panconesi. *Rumour spreading and graph conductance*, pages 1657–1663.

[13] Milena Mihail, Christos Papadimitriou, and Amin Saberi. On certain connectivity properties of the internet topology. *Journal of Computer and System Sciences*, 72(2):239 – 251, 2006. JCSS FOCS 2003 Special Issue.

[14] Daniel A. Spielman and Shang-Hua Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing*, STOC '04, pages 81–90, New York, NY, USA, 2004. ACM.

[15] Flavio Chierichetti, Silvio Lattanzi, and Alessandro Panconesi. Almost tight bounds for rumour spreading with conductance. In *Proceedings of the Forty-second ACM Symposium on Theory of Computing*, STOC '10, pages 399–408, New York, NY, USA, 2010. ACM.

[16] Keren Censor-Hillel and Hadas Shachnai. Fast information spreading in graphs with large weak conductance. *SIAM Journal on Computing*, 41(6):1451–1465, 2012.

[17] Keren Censor Hillel and Hadas Shachnai. Partial information spreading with application to distributed maximum coverage. In *Proceedings of the 29th ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing*, PODC '10, pages 161–170, New York, NY, USA, 2010. ACM.

[18] A. Sinclair. In *Algorithms for Random Generation and Counting: A Markov Chain Approach*. Birkhäuser Basel, 1993.

[19] Burcu Sayin and Serap Şahin. A novel approach to information spreading models for social networks. In *Proceedings of the 6th International Conference on Data Analytics*, pages 23–27. IARIA, 2017.

[20] Yuanyuan Bao, Chengqi Yi, Yibo Xue, and Yingfei Dong. A new rumor propagation model and control strategy on social networks. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 1472–1473, New York, NY, USA, 2013. ACM.

[21] Emilio Serrano, Carlos Ángel Iglesias, and Mercedes Garijo. A novel agent-based rumor spreading model in twitter. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 811–814, New York, NY, USA, 2015. ACM.

[22] Gennaro Cordasco, Luisa Gargano, Adele A. Rescigno, and Ugo Vaccaro. Brief announcement: Active information spread in networks. In *Proceedings of the 2016 ACM Symposium on Principles of Distributed Computing*, PODC '16, pages 435–437, New York, NY, USA, 2016. ACM.

[23] Juan Zhang, Jianquan Li, and Zhien Ma. Global dynamics of an seir epidemic model with immigration of different compartments* *this research is supported by the nnsf of china (19971066). *Acta Mathematica Scientia*, 26(3):551 – 567, 2006.

[24] Sumith N, Annappa B, and S. Bhattacharya. Rnsir: A new model of information spread in online social networks. In *2016 IEEE Region 10 Conference (TENCON)*, pages 2224–2227, Nov 2016.

[25] Chao Tong, Wenbo He, Jianwei Niu, and Zhongyu Xie. A novel information cascade model in online social networks. *Physica A: Statistical Mechanics and its Applications*, 444(Supplement C):297 – 310, 2016.

[26] Networkx package webpage. https://networkx.github.io/.

[27] David Kempe, Alin Dobra, and Johannes Gehrke. Gossip-based computation of aggregate information. In *Proceedings of the 44th Annual IEEE Symposium on Founda-*

tions of Computer Science, FOCS '03, pages 482–, Washington, DC, USA, 2003. IEEE Computer Society.

[28] Chuang Liu and Zi-Ke Zhang. Information spreading on dynamic social networks. 19:896–904, 04 2014.

[29] Danah Boyd and Eszter Hargittai. Facebook privacy settings: Who cares? 15, 07 2010.

[30] Debarati Gangopadhyay, Saswati; Dhar. Social networking sites and privacy issues concerning youths. 5:1, June 2014.

[31] Matt McKeon. The evolution of privacy on facebook, 2010. http://mattmckeon.com/facebook-privacy/.

[32] John Edens. Facebook privacy policy has become less transparent, harder to understand and control, experts say, March 2017. http://www.stuff.co.nz/technology/social-networking/89645571/Facebook-privacy-policy-has-become-less-transparent-harder-to-understand-and-control-experts-say.

[33] Jennifer Shore and Jill Steinman. Did you really agree to that? the evolution of facebook's privacy policy, August 2015. https://techscience.org/a/2015081102/.

[34] A. Ho, A. Maiga, and E. Aimeur. Privacy protection issues in social networking sites. In *2009 IEEE/ACS International Conference on Computer Systems and Applications*, pages 271–278, May 2009.

[35] Virpi Tuunainen, Olli Pitkänen, and Marjaana Hovi. *Users' Awareness of Privacy on Online Social Networking Sites Case Facebook*. 2009.

[36] N. Talukder, M. Ouzzani, A. K. Elmagarmid, H. Elmeleegy, and M. Yakout. Privometer: Privacy protection in social networks. In *2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*, pages 266–269, March 2010.

[37] Elena Zheleva and Lise Getoor. To join or not to join: The illusion of privacy in social

networks with mixed public and private user profiles. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 531–540, New York, NY, USA, 2009. ACM.

[38] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P.N. Puttaswamy, and Ben Y. Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European Conference on Computer Systems*, EuroSys '09, pages 205–218, New York, NY, USA, 2009. ACM.

[39] Christo Wilson, Alessandra Sala, Krishna P. N. Puttaswamy, and Ben Y. Zhao. Beyond social graphs: User interactions in online social networks and their implications. *ACM Trans. Web*, 6(4):17:1–17:31, November 2012.

[40] Tabitha L. James, Merrill Warkentin, and Stéphane E. Collignon. A dual privacy decision model for online social networks. *Inf. Manage.*, 52(8):893–908, December 2015.

[41] Cailing Dong, Hongxia Jin, and Bart Knijnenburg. Predicting privacy behavior on online social networks. 2015.

[42] Cailing Dong, Hongxia Jin, and Bart P. Knijnenburg. Ppm: A privacy prediction model for online social networks. pages 400–420, 2016.

[43] E. Aghasian, S. Garg, L. Gao, S. Yu, and J. Montgomery. Scoring users' privacy disclosure across multiple online social networks. *IEEE Access*, 5:13118–13130, 2017.

[44] A. Srivastava and G. Geethakumari. Measuring privacy leaks in online social networks. In *2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 2095–2100, Aug 2013.

[45] Di Wang, Xiao-Jun Zeng, and John A. Keane. A simplified structure evolving method for mamdani fuzzy system identification and its application to high-dimensional problems. *Information Sciences*, 220(Supplement C):110 – 123, 2013. Online Fuzzy Machine Learning and Data Mining.

[46] Facebook privacy basics, 2017. https://tr-tr.facebook.com/about/basics.

[47] Proto.io website, which was used to create the mockup, 2017. https://proto.io/.

[48] Access link of the mockup implementation video, 2017. https://www.dropbox.com/s/tmnjxf46r225rzq/mockup.mp4?dl=0.

APPENDIX A

MAIN SCREENS OF THE MOCKUP

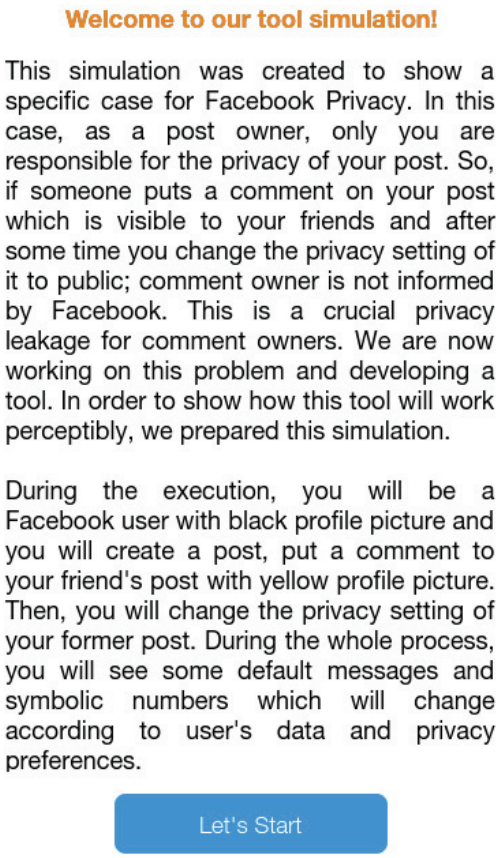

Figure A.1. Main Screen 1

Figure A.2. Main Screen 2

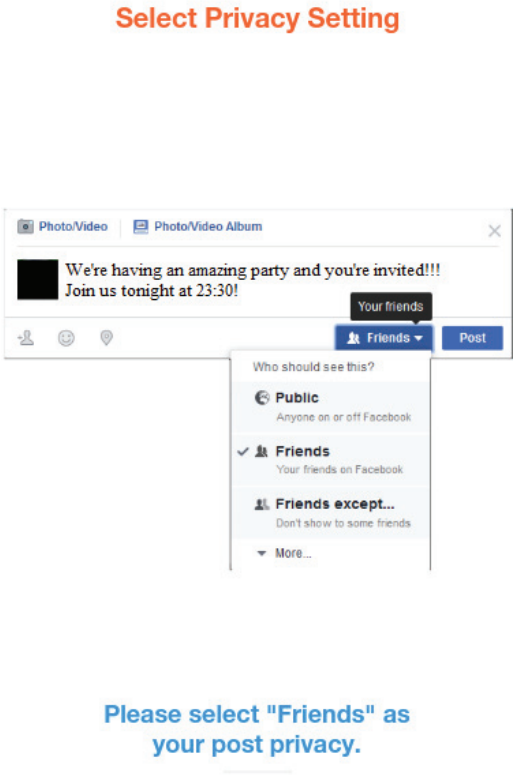

Figure A.3. Main Screen 3

Figure A.4. Main Screen 4

Figure A.5. Main Screen 5

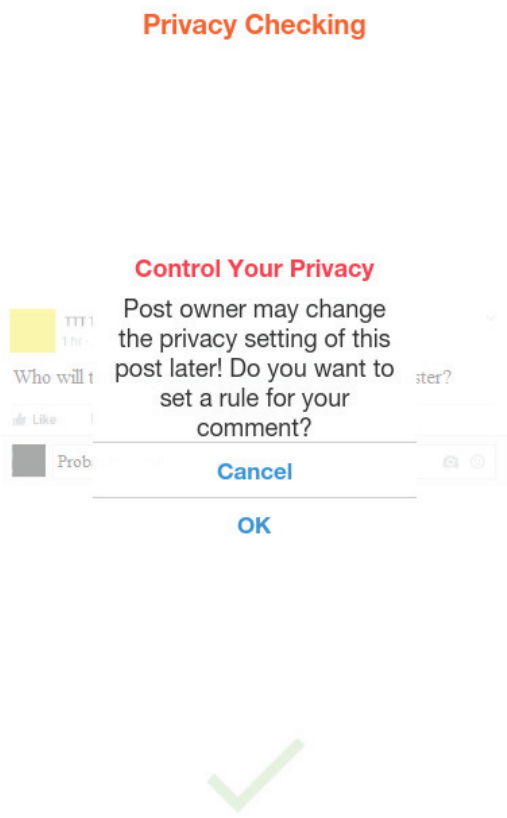

Figure A.6. Main Screen 6

Figure A.7. Main Screen 7

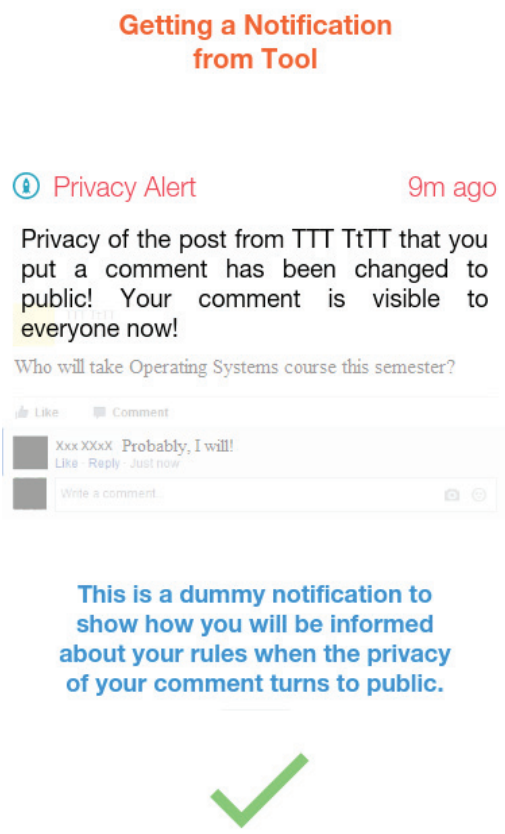

Figure A.8. Main Screen 8