Simulating the Effect of Privacy Concerns in Online Social Networks

Nathalie Baracaldo, Claudia López, Mohd Anwar, Michael Lewis
Email: [nab62,cal95, manwar] @pitt.edu, ml@sis.pitt.edu
School of Information Sciences
University of Pittsburgh, Pittsburgh, USA

Abstract

While sharing information is the main purpose of online social networks (OSN), privacy is a major concern. It has been argued that the lack of privacy protection caused the massive departure of users from MySpace in 2008 [1] – specially, due to mass media coverage on dangers of teenagers' public profiles and the appearance of Facebook, which provided a more private OSN. Besides mass media, we explore the impact of family and friends in instilling privacy concerns on teenagers. We develop an agent-based model to simulate propagation of privacy concerns through connection in a social network, and the effect of individual users' privacy decisions on the level of overall participation in an OSN. We report the results of simulating different privacy propagation factors. Our results are useful to understand and predict hidden patterns of user behavior that lead to changes in the number of active users in an OSN.

1 Introduction

In November of 2006, MySpace was the largest Online Social Network (OSN) in the world, with more than 130 million users [2]. After having this large amount of users, a massive departure from MySpace took place. The reasons that led to this event are not clear, although several studies have tried to identify the main factors that contributed to this phenomenon. Among these factors is the lack of privacy mechanisms to protect the information of users. In that particular period of time, the great majority of users of MySpace were teenagers who inadvertently disclosed their private information.The exposure of teenagers' information led to incidents in which teenagers were victims of sexual predators [3]. The media repeatedly reported cases in which teenagers' safety was threatened due to their participation in MySpace. This may have led to the concern of parents and schools to educate and make teenagers leave MySpace. Hence, both mass media and family pressure seem to be plausible factors that influenced the departure.

Another possible factor that several studies point out as the main driver to the diaspora of users from MySpace was the appearance of Facebook. According to this hypothesis, Facebook offered a more private space for teenagers to interact with their peers than that of a more open OSN such as MySpace.

In this paper, we look for the most plausible factors that lead to the massive departure from MySpace. To answer this, we propose a model to simulate how different factors modify the user base of an OSN. Based on an extensive literature review, we included in our model the following factors *media, peer* and *family pressure. Media pressure* refers to the impact of media coverage about privacy breaches on raising individual's privacy concerns, which we hypothesize to have impacted some users' decisions on ceasing their participation in MySpace. *Peer pressure* is defined as the influence of peers on one's decision to stay or leave an OSN, where the reason to leave might be joining another OSN or stoping using OSNs. Since we are considering teenagers as one of the relevant groups that is part of any OSN, and in particular of MySpace, we also include *family pressure* as a factor that affects teenagers' awareness to privacy and that possibly influenced them to leave MySpace. Since media brings consciousness of privacy threats to families, we consider family pressure as a function of media pressure. The contributions of our research are as follows:

1. Extrapolating from literature review, we identify the main factors that may have raised privacy awareness, which result in the massive departure from MySpace.

2. We design an agent-based model to simulate and analyze how individual privacy concerns and environmental variables affect the global community behavior in an OSN.

3. We develop a generic model to capture different factors that influence the behavior of users of any OSN. As a result, this model may help to predict and improve the sustainability of OSNs.

The rest of the paper is organized as follows. In Sec-

tion 2, we describe agent-based modeling methodology and basic properties of OSNs. In Section 3, we present the proposed model. In Section 4, we present the results. In Section 5, we conclude and present future work.

2 Background

Agent-Based Modeling: Agent-based Modeling has been successfully used in social psychology to identify how individuals' decisions lead to a global phenomenon and predict outcomes of actions in a particular environment over time. This methodology has been applied to many domains such as innovation networks [5] and partner selection patterns [6]. Recently it has been also used to find the critical mass of OSNs [7] and designing successful OSNs [8].

In an agent-based model, each individual, known as an agent, acts based on their perceived environment (e.g., what other agents do) and a set of rules representing their programmed behaviors. Each agent has its own behavior, which is believed to be part of the underlying factors to create a pattern in a society. The advantages of this methodology are the ability to predict global behavior patterns based on individuals' behavior without the cost of creating the real conditions to test potential factors causing them. An understanding of these patterns serves to anticipate the repercussions of changes of environmental variables. In this paper, we use agent-based modeling to better understand the drivers that resulted in the massive departure from MySpace.

Topology of OSNs: OSNs are generally modeled as an undirected graph in which a node represents a user and a link between a pair of nodes represents a friend relationship. The number of friends a user has is known as degree of the corresponding node. For large OSNs, the *degree distribution*, which is the distribution of friendship links in a network, follows a power-law distribution: $P(k) \sim k^{-\gamma}$, where k is the node degree and $\gamma \leq 3$, meaning that very small number of nodes have a very large number of links [2]. OSNs also follow the *small-world* property, in which networks have a small diameter and high clustering [9]. The *assortativity* of a graph is a measure of the likelihood for nodes to connect to other nodes with similar properties.

MySpace Topology: To simulate the propagation of privacy concerns in a way that reflects reality, we collected available data about MySpace's topology. A method to estimate MySpace's topology measures was reported in [2]. Table 1 shows the sampling information and estimated metrics for MySpace. The estimated mean number of friends was 137.1 and the estimated average degree of separation was 2.7. The estimated average clustering coefficient and assortativity measure show that MySpace was not a fully connected graph, but a graph with a tendency to form clusters with hubs connecting nodes with smaller degree. The

Table 1. MySpace Topology Metrics [2]

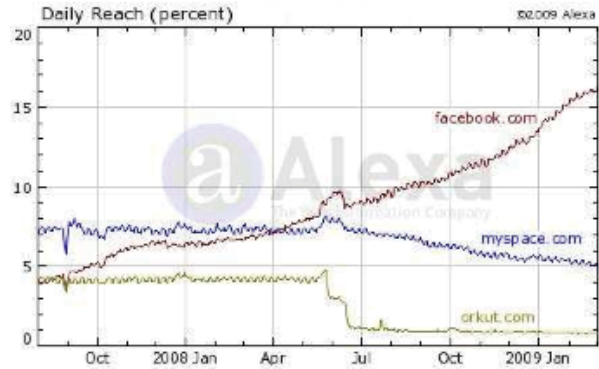| Sampling ratio	0.08%
Number of nodes	100,000
Number of edges	6,854,231
Mean degree (k)	137.1
Average clustering coefficient	0.26
Assortativity	0.02
Estimated degree of separation	2.7
Estimated γ	2.7 - 3.1

Figure 1. Daily Accesses to MySpace [1]

value of γ indicates that while a very small number of users had many connections, the great majority of users had very few.

Torkjazi *et al.* [1] reported a detailed analysis of 360,000 MySpace profiles. Their results show that the number of inactive accounts grew exponentially since 2006 to the middle of 2007, while the number of active accounts grew linearly. From 2007 to the middle of 2008, both curves are linear, with the number of inactive accounts having a higher slope.

The decline of MySpace was also documented by Alexa.com, which reports the number of daily unique visitors to websites. Figure 1 illustrates the change in the amount of people visiting MySpace. We use the information about the user participation in MySpace to validate our model.

3 Model Specification

In this section, we present our model, which identifies the most significant factors that led to the massive departure of users, especially teenagers. We begin by explaining the following factors: *peer, media* and *family pressure*.

Peer Pressure: Peer pressure occurs "when people your own age encourage you to do something or to keep from doing something else, no matter if you personally want to

or not" [4]. The most relevant aspect of this definition is the fact that "individuals are motivated to act and think in certain ways because they have been urged, encouraged, or pressured by a peer to do so" [10]. Several studies suggest that people are subject to peer pressure when joining and leaving an OSN. It has been shown that the probability to enter an OSN increases if the amount of friends that a person already has in that community surpasses a threshold [11]. Conversely, it has been stated that people left MySpace in groups, not individually [1]. From the privacy perspective, peer pressure is similar to the *herd principle* [12], also known as *social proof* [13], in which individuals tend to accept taking risks if the people surrounding them appear to be taking the risk too.

Teenagers are particularly vulnerable to peer pressure. They find a space in OSNs to develop the important skills of adapting to membership in a social group and, as such, feel the pressure to join those communities or move to another ones when their peers do so. For these reasons and evidence found in [1, 11], we include peer pressure as a factor that may influence the decline of user participation in an OSN.

Media Pressure: It has been shown that mass media is responsible for establishing the topics of discussion in a community at a particular moment in time [15]. According to [16], in a year, on average a person spends 40% of the time using mass media, comparing to 33% of time spent sleeping, and 26% of the time performing other activities. Due to the enormous exposure of people to media, we believe that mass media plays an important role on the thoughts parents and teenager have related to privacy. In the period of interest (2005 to 2009), mass media presented several news related to privacy problems, damage of reputation and the risks of publishing teenagers information on OSN. In particular, concerns arose due to the use of teenagers' information by sexual predators. Figure 2 presents the number of news related to MySpace's privacy issues during the period elapsed between January 2005 to December 2009. Initially the media did not have any particular interest on the topic, but around January 2006 a series of privacy related news started to surface. We believe that this media coverage motivated families and the overall community to persuade teenagers to leave MySpace.

Teenagers and Family Pressure: According to [14], a significant proportion of MySpace users were teenagers between 14 and 21 years old. Teenagers are exposed to family pressure as documented in [17], where teens that do not use MySpace are classified into two classes: 1) *disenfranchised teens* and 2) *conscientious objectors*. Disenfranchised teens are teenagers that cannot access Internet for non-academic purposes due to parental restrictions or because they only have access to Internet in the school. The group of conscientious objectors is conformed by teens that agreed with their parents not to disclose information in OSNs due to moral or
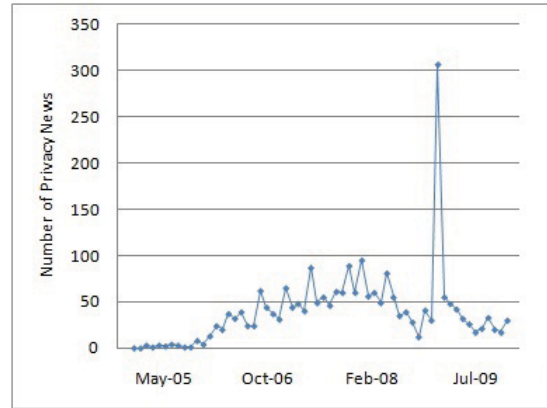

Figure 2. Number of news articles related to privacy issues on MySpace (2005 to 2009)

safety concerns. We define *family pressure* as the pressure exerted by family members on teenagers to avoid the disclosure of information that may later endanger or damage their reputation.

3.1 Mathematical Formulation

We use agents to represent users in an OSN. In our model, agents are exposed to the influence of their peers, mass media and family. These pressures are modeled as continuous functions in the interval [0–1]. When a pressure equals to 1, it means that an agent is experiencing the maximum pressure possible, and a pressure of 0 means that the agent is not experiencing any pressure at all. We denote $\rho(t)$, $\mu(t)$ and $\Upsilon(t)$ as peer, media and family pressure at a time instant t, respectively.

In order to model the pressure experienced by agents, we consider that privacy concerns which result in staying or leaving an OSN spread like a virus[1]. The virus spread model starts with an initial *breakout*, which is the number of initial infected people. The probability of an individual getting infected increases with the number of infected neighbors. Hence, the virus continuously spreads through the neighbors. We adapted the model presented in [19], in which a technology spreads like a virus and the neighbors' status influences the decisions made by an agent to adopt a new technology.

Each agent is either *active* or *inactive*. An inactive agent represents a user that has left the OSN, while an active agent is one that still belongs to the OSN (using the virus metaphor, an active friend is healthy and an inactive agent has caught the virus). The number of friends of an agent is denoted by $\mathcal{K}$ (number of active plus number of inactive

---

[1]Ideas were shown to spread like viruses in [18].

friends) and the number of friends that are inactive is represented by $\mathcal{P}$.

We use two alternative functions to model peer pressure: *fractional peer pressure* $\rho_f(t)$ and *linear peer pressure* $\rho_l(t)$. We use $\rho(t)$ to refer to both of them when we do not want to distinguish. These alternative peer pressure functions are defined as follows:

$$\rho_f(t) = \left(\frac{\mathcal{B}}{\mathcal{C}}\right) * \left(\frac{\mathcal{P}}{\mathcal{K}}\right); \quad \rho_l(t) = \left(\frac{\mathcal{B}}{\mathcal{C}}\right) * (\mathcal{P})$$

where $\mathcal{B}$ represents the benefit of leaving the OSN and $\mathcal{C}$ represents the cost of leaving the OSN, here $0 \leq \frac{\mathcal{B}}{\mathcal{C}} \leq 1$, and, the ratio $\frac{\mathcal{P}}{\mathcal{K}}$ represents peer pressure that an agent feels to leave or stay in the OSN. Intuitively, the cost of departing from an OSN includes losing access to its information. In addition, when a new OSN, such as Facebook, appears with additional positive factors (e.g., more privacy), the benefit of leaving the initial OSN increases as more friends decide to join the new OSN (i.e., Facebook). That is, as the number of friends that leaves the OSN increases, the peer pressure to leave increases as well.

Media pressure is modeled as a function that considers the normalized amount of privacy related news at a particular moment of time. In Section 3.2, we present the different functions used for the simulations.

We modeled *family pressure* as a consequence of media pressure. We believe that the media coverage takes some time to raise concerns among family members about privacy issues for teenagers. We call this period of time a *lag*. At a time instant t, the family pressure is given by a previous value of the media pressure as follows: $\Upsilon(t) = \mu(t - i)$, where i represents the magnitude of the *lag* period.

We define the *total pressure* (Γ) as a function of time composed by peer, media and family pressure as follows:

$$\Gamma(t) = \begin{cases} \rho(t) * (w_1 \mu(t) + w_2 \Upsilon(t)) & \text{if } w_1 + w_2 > 0 \wedge \rho \neq 0 \\ \rho(t) & \text{if } w_1 + w_2 = 0 \\ w_1 \mu(t) + w_2 \Upsilon(t) & \text{if } \rho(t) = 0 \end{cases}$$

Where, $0 \leq w_1 \leq 1$ and $0 \leq w_2 \leq 1$ are weights that represent the influence of media and family pressure. When an agent represents an adult, the weight associated to the family pressure is always zero ($w_2 = 0$) to reflect the fact that adults are not subject to family pressure.

Finally, each agent is randomly assigned a threshold (τ) that determines whether environmental pressures would make her leave the OSN. We model τ as a normally distributed variable in the range ($0 \leq \tau \leq 1$). An agent leaves the OSN when the total pressure is greater than its threshold ($\Gamma(t) \geq \tau$). Hence, τ determines the sensibility of an agent.

3.2 Considered Scenarios

We use different scenarios to observe how media pressure influences the simulation.

1) Constant Media Pressure: This scenario simulates periods of time during which media privacy coverage is constant. In this case, the media pressure is set to a given number $0 \leq c \leq 1$, $\mu(t) = c$ for all instance of time t.

2) Random Sinusoidal Pressure: In order to simulate a scenario in which the media pressure oscillates as the time goes by, we evaluate the system under a media pressure given by the following equation: $\mu(t) = |\ sin(t) * random\ |$, where, in order to introduce some randomness in the sinusoidal function, we multiply it by a uniformly distributed random number. The absolute value is taken to ensure that the $\mu(t)$ is always positive.

3) Observed Media Pressure: We collected the number of news related to privacy issues on MySpace during the period of January 2005 to December 2009, shown in Figure 2. We used this information to run simulations with media pressure as a linear function of the number of news.

4) No media pressure: This option allows us to run the simulation without considering media influence on raising privacy awareness.

4 Results

We implemented our model using NetLogo 4.1.2 environment [20]. Figure 3 shows the interface of our implementation. The results reported in this paper came from approximately 13,000 simulations that were run using a scale-free network of 1000 nodes. The average cluster coefficient of the network is 0.31, average degree is 5.056 and the network diameter is 5. The input parameters of the network generation algorithm were specified in order to get similar network measures to those estimated for MySpace (see Table 1). The proportion of teenagers to the total number of users was kept to 0.75 to reflect the configuration of MySpace [14]. Each agent was randomly assigned a privacy threshold τ, which was normally distributed with mean of 0.5 and standard deviation of 0.15.

The parameters that were tested in the simulations are the following.

- Media Pressure: Constant, Sinusoidal, Observed, None (see Section 3.2)
- Peer Pressure: None, fractional ($\rho_f(t)$), linear ($\rho_l(t)$)
- Ratio $\mathcal{B}/\mathcal{C}$: 0.16, 0.33, 0.5, and 1
- Initial outbreak sizes: 3%, 6%, and 10% of the population
- Family pressure: with and without family pressure
- Lag: 10 days, 30 days, and 50 days
- Media factor (w_1): 0.3 and 0.5
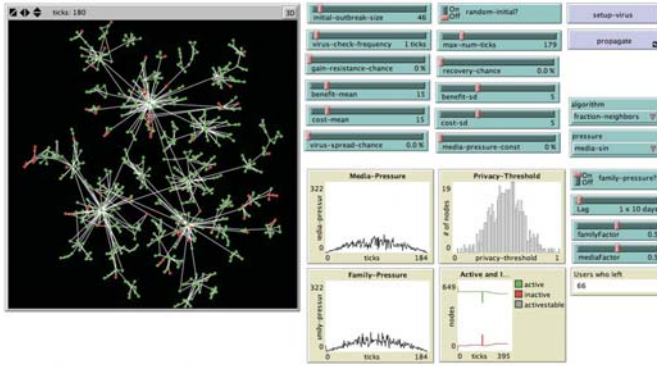- Family factor (w_2): 0.7 and 0.5

Figure 3. A screenshot of a simulation

- Periods of time: 500, 750, 1000, 1500 and 1790 days

The simulations were run for all the combinations of the parameters previously mentioned. The maximum interval of time in which the simulations were run is equivalent to the period between January 2005 and December 2009. Each agent evaluated if she wanted to leave the OSN every 30 days or less. Agents were not allowed to return to the OSN once they left it.

In order to analyze the data generated by the 13,000 simulations, we use the C4.5 classification algorithm [21] [2]. This algorithm allows us to find the variables that better predict the amount of people leaving the community (e.g., type of media pressure, kind of peer pressure). In order to use this classification algorithm, the number of inactive agents at the end of each simulation was categorized in intervals of 100 people (e.g. [0–99], [100–199], and so on). We run the algorithm with a confidence factor of 0.25 and 100 as a minimum number of instances per class.

Our simulation results suggest that mass media and peer pressure play the most relevant roles when predicting the number of people leaving an OSN. We analyzed the results according to the scenarios presented in Section 3.2.

Constant Media Scenario: For this scenario, the number of people leaving the OSN depends on the constant value used as media pressure and the presence peer pressure. A high constant value of the media can lead to every agent leaving, and a low constant value can lead to the departure of only the initial infected agents. The final number of inactive agents is reached very quickly in the first iterations and then is kept constant until the end of the simulation. This scenario does not reflect MySpace experience, as it does not produce the smooth decreasing trend seen in Figure 1. These results suggest that peer pressure was not the unique cause of the massive departure from MySpace.

No-Media Scenario: When media pressure is not considered, some agents still leave the community. In this case,

the initial outbreak size and peer pressure play the most important role. The rate of departure increases when $\mathcal{B}/\mathcal{C}$ ratio is higher. This scenario shows similar results to the ones found for *constant media* scenario, where the final number of inactive users is reached earlier on the simulation (although, it is slower than in constant media scenario), and then it is kept constant.

Random Sinusoidal Media Scenario: When we consider the sinusoidal function as media pressure, the number of agents leaving the OSN is always more than 40%. For this scenario, the most important factor is the existence of family pressure. If there is no family pressure, the number of inactive agents range from 500 to all of them. This difference is explained by the number of days considered in the simulations. When the simulation does not include higher values of the first increasing sinus curve, about half of the agents leave. In contrast, when the simulation runs for long periods (when the media pressure reaches values closer to 1), everyone leaves. When family pressure is present, the number of days included in the simulation is also an important factor. The pattern is the same –the longer the period, the higher the number of users quitting the OSN. Other influencing factors are the kind of peer pressure, the ratio $\mathcal{B}/\mathcal{C}$ and family factor. In general, the *fractional peer pressure* function produces more moderated values of inactive people (from 40% to 85%) compared to *linear peer pressure* (50% to 95%). For these reasons, the random sinusoidal media scenario does not resemble the real world pattern shown in Figure 1.

Observed Media Scenario: When observed media pressure is used, the number of days, the initial outbreak size and the type of peer pressure are the most influencing factors. The longer the period of analysis, the higher the number of people leaving. For shorter periods (less than 750 days), the outbreak size and the kind of peer pressure are key. In this case, the final number of people leaving is always less than 30%. For longer intervals, the main factor is the existence of peaks on media-pressure. Considering the peak on the number of news reported in January 2009, almost everyone leaves the OSN. In contrast, when family pressure weight (w_2) is higher than the media pressure weight (w_1) and the family pressure *lag* is higher than 30 days, the results differ from the latter scenario. In this case, the effect of a high media pressure is diminished by a low family pressure. This finding suggests that in reality the influence of family on teenagers is relevant. On the other hand, when media peaks are not considered, the *lag* of the family pressure and peer pressure are the key. If there is any lag, the existence of peer pressure leads to the departure of almost half of people. When the lag is longer, the amount of people leaving ranges from 20% to 50%. The highest values of inactive agents occur when the ratio $\mathcal{B}/\mathcal{C}$ is higher than 0.5. Thus, when the cost of leaving is closer to the

value of the benefit, people are more likely to leave. The simulations that use the *fractional peer pressure* function generated more moderated departure, ranging from 20% to 30 %. The results achieved in this scenario were the closest to the real data collected about MySpace. The results found when the media peak is not included match the curve shown in Figure 1. Further analysis needs to be performed to understand how effects of peaks on media coverage are diminished in real life.

Although we run our simulations with a small network, we were able to identify some interesting patterns under different conditions of media, peer and family pressure. Further simulations are required to generalize these results. We aim to run new simulations using bigger networks with other measures in order to analyze potential changes on the identified patterns.

5 Conclusions and Future Work

In this paper we present an agent-based model to identify the main factors that contributed to the massive departure of MySpace in 2008. This model provides a framework that can be extended to test additional hypothesis about the causes that may lead to massive departures from any OSN.

The presented model includes three influencing factors on the decision of leaving or staying in an OSN: peers, mass media and family. We simulated 4 different scenarios for media pressure: no-media, constant, random sinusoidal and observed media pressure. We run more than 13,000 simulations and analyzed the data using a classification algorithm. The results suggest that the media and peer pressure are the most influencing factors. We found that the model using observed media coverage produces curves of number of inactive users over time similar to MySpace historic data. However, this finding is challenged by the existence of a peak on number of news related to privacy in this OSN. Our insight is that considering media pressure as a linear function of news might be still too simple to grasp the complexities of how media affects people's decisions. As future work, we plan to evaluate additional mechanisms to estimate media pressure. It would also be interesting to analyze the impact of additional factors such as gender, privacy legislation and privacy policies offered by the OSN provider.

References

[1] M. Torkjazi, R. Rejaie, and W. Willinger, "Hot today, gone tomorrow: on the migration of myspace users," in *Proc. of the 2nd ACM workshop on Online social networks*, 2009, pp. 43–48.

[2] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *Proc. of the 16th international conference on World Wide Web*, 2007, pp. 835–844.

[3] S. B. Barnes, "A privacy paradox: Social networking in the United States," *First Monday*, vol. 11, no. 9, Sep. 2006.

[4] B. B. Brown, D. R. Clasen, and S. A. Eicher, "Perceptions of peer pressure, peer conformity dispositions, and self-reported behavior among adolescents," *Developmental Psychology*, vol. 22, no. 4, pp. 521 – 530, 1986.

[5] N. Gilbert, A. Pyka, and P. Ahrweiler, "Gilbert, pyka and ahrweiler: Innovation networks- a simulation approach," *Journal of Artificial Societies and Social Simulation*, pp. 1–13, 2001.

[6] S. M. Kalick and T. E. Hamilton, "The matching hypothesis reexamined," *Journal of Personality and Social Psychology*, vol. 51, no. 4, pp. 673–682, 1986.

[7] W. Wiersma, "Thesis proposal - Simulating the Emergence of Critical Mass in Online Communities," 2011.

[8] Y. Ren and R. E. Kraut, "A simulation for designing online community: Member motivation, contribution, and discussion moderation," 2011.

[9] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proc. of the 7th ACM SIGCOMM conference on Internet measurement*, 2007 , pp. 29–42.

[10] D. A. Santor, D. Messervey, and V. Kusumakar, "Measuring Peer Pressure, Popularity, and Conformity in Adolescent Boys and Girls: Predicting School Performance, Sexual Attitudes, and Substance Abuse," *Journal of Youth and Adolescence*, vol. 29, no. 2, pp. 163–182, 2000.

[11] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan, "Group formation in large social networks: membership, growth, and evolution," in *Proc. of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006

[12] F. Stajano and P. Wilson, "Understanding scam victims: seven principles for systems security," *Commun. ACM*, vol. 54, pp. 70–75, 2011.

[13] R. Cialdini, *Influence: Science and Practice*. Pearson Education, 2009.

[14] M. Thelwall, "Social networks, gender, and friending: An analysis of myspace member profiles," *J. Am. Soc. Inf. Sci. Technol.*, vol. 59, pp. 1321–1330, June 2008.

[15] J. Zaller, *The Mith of Massive Media Impact Revived: New support for a Discredited Idea*, 1996, ch. 2 The Myth of Massive Media Impact Revived: New Support for a Discredited Idea.

[16] S. Biagi, *Media/Impact:An Introduction to Mass Media*, 2006, ch. 13 Society, Culture and Politics: Shaping the issues.

[17] danah boyd, "Why youth (heart) social network sites: The role of networked publics in teenage social life," *MacArthur Foundation Series on Digital Learning*, vol. Youth, Identity, and Digital Media Volume, pp. 119–142.

[18] S. Godin, *Unleashing the Ideavirus*. Hyperion, 2001.

[19] D. López-Pintado, "Diffusion in Complex Social Networks," *Information Sciences*, pp. 1–28, 2004.

[20] U. Wilensky, "Netlogo," Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL., Tech. Rep.

[21] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. , 1993.