Spam diffusion in a social network initiated by hacked e-mail accounts

Ghita Mezzour

Department of Electrical and Computer Engineering,
Carnegie Mellon University,
Pittsburgh 15213, PA, USA
E-mail: mezzour@cmu.edu

Kathleen M. Carley

School of Computer Science,
Carnegie Mellon University,
Pittsburgh 15213, PA, USA
E-mail: kathleen.carley@cs.cmu.edu

Abstract: Rumour diffusion has been widely studied in the literature. However, most prior work assumes that the rumour initiators are humans. This assumption does not hold in cyber space where some of the accounts are hacked. These hacked accounts aggressively disseminate rumours by continuously sending a large amount of messages. Our results show that when rumours are initiated by hacked accounts, the rumour diffusion dynamics are different. This work suggests that social simulations of interactions over cyber space should capture the fact that spamming accounts also participate in these interactions. The presence of these accounts can alter the simulation dynamics.

Keywords: attack propagation; spam diffusion; social networks; attack modelling; rumour diffusion; simulation; agent-based modeling; agent-based simulation; security; hacked accounts.

Reference to this paper should be made as follows: Mezzour, G. and Carley, K.M. (xxxx) 'Spam diffusion in a social network initiated by hacked e-mail accounts', *International Journal of Security and Networks*, Vol. x, No. x, pp.xx–xx.

Biographical notes: Ghita Mezzour is a PhD Student in the Electrical and Computer Engineering at Carnegie Mellon University. Her research typically combines cyber-security and social networks. She is interested in addressing cyber-security and privacy issues in social networks. Another important component of her research is the global cyber threat assessment. In that line of research, she uses multiples cyber security data sets and applies social network analysis techniques. She holds a Master and a Bachelor in Communication Systems from the Swiss Federal Institute of Technology in Lausanne.

Kathleen M. Carley, Harvard PhD, is a Professor of Computation, Organisations and Society in the School of Computer Science at Carnegie Mellon University and the Director of the Center for Computational Analysis of Social and Organisational Systems (CASOS). Her research areas include social network analysis, dynamic network analysis, agent-based modelling, computational social and organisation theory, adaptation and evolution, social network text mining, cyber security, information diffusion, social media and telecommunication, disease contagion, and disaster response. She and members of her center have developed novel tools and technologies for analysing large scale geo-temporal multi-dimensional networks (ORA) and agent-based simulations (Construct).

1 Introduction

Spam diffusion over social networks is a major problem. Spam content varies between counterfeit watches, fake commercial offers and erroneous political information. In some instances, people think that the spam content is correct and disseminate it to their contacts. This was, for example, observed in the political discussion on Twitter about the recent Massachusetts senate race (Metaxas and Mustafaraj, 2012; Ratkiewicz et al., 2011). Spamming accounts disseminated false information and legitimate users thought the information is correct and disseminated it. Spamming accounts are either fake accounts or hacked accounts. Fake accounts are accounts created for the only purpose of sending spam or rumours. On the other hand, hacked accounts belong to legitimate users, but a hacker has gained access to these accounts. Hackers gain access to accounts by guessing passwords, stealing password databases, or through a computer malware. Account hacking is a very prevalent cyber-attack. Multiple online websites and videos explain how to compromise e-mail accounts (Go Hacking, 2008; Hacker The Dude, 2013). Facebook reveals that it daily detects 600,000 attempts to compromise accounts (ConsumerReports.org, 2011). Spam sent from hacked accounts is more credible. For example, when the hacked Twitter account of Fox news announced that Obama was shot dead in July 2011, the information rapidly spread out in the Internet (Guardian, 2011). Similarly, when the hacked Twitter account of the presidential adviser for disaster management posted a false tsunami warning in 2011, the entire nation was scared (Yahoo! News, 2010).

Information diffusion in social networks is widely studied (Zanette, 2002; Wu et al., 2003; Carley, 1991; Valente, 1996; Granovetter, 1987; Goldenberg, 2001; Domingos and Richardson, 2001; Kempe et al., 2003; Morris, 2000). However, most prior work overlooks the case where compromised accounts aggressively disseminate the information. In this paper, we modify a traditional information diffusion model (Carley, 1991) in order to capture the behaviour of hacked accounts. More specifically, the modified model has two types of agents: hacked agents that represent hacked accounts and regular agents that represent people that have no hacked account. Hacked agents continuously transmit a rumour to all their contacts, whereas regular agents transmit the rumour at a lower rate and may lose interest in transmitting the rumour. Our results show that rumour diffusion dynamics are very different in the model that accounts for the behaviour of hacked accounts and the model that does not. More specifically, when the behaviour of hacked accounts is accounted for, spam diffusion is faster and reaches more people. Moreover, parameters like the social network size affect differently the diffusion in two models.

We discuss related work in Section 2. We provide background on Construct (Carley, 1991), the agent-based modelling tool we use and the Box–Behnken experiment design in Section 3. We present our model in Section 4, our virtual experiment in Section 5 and results in Section 6. We discuss limitations and future work in Section 7 before concluding.

2 Related work

Diffusion through 'word-of-mouth' has been widely studied in the literature in different contexts using a variety of approaches. However, most of this prior work does not consider the case where some of the participants in the diffusion are hacked accounts that behave maliciously. Much previous work considers that information propagates similarly to diseases (Zanette, 2002; Wu et al., 2003). Examples of classical disease propagation models are SIR and SIRS, which are based on the stages of a disease in a person. Initially, people are susceptible (S) to the disease. When a person gets the disease, they become infected and infectious (I). When a person recovers, they become recovered (R). A recovered person is immune against the disease. In the SIRS model, a person can become susceptible (S) again. Using epidemic models allows benefiting from tremendous work on epidemic diffusion. However, information is intrinsically and biologically different from diseases. Construct (Carley, 1991) is a powerful turn-based and agent-based modelling tool for information and belief diffusion. Agents have information, beliefs and transactive memory. Each time period, agents are paired and can exchange information, belief and transactive memory. Construct is the tool we use in this paper and we cover it in more detail in Section 3. Diffusion of innovation (Valente, 1996; Granovetter, 1987; Goldenberg, 2001) has been studied through threshold models and cascade models. In the threshold model (Granovetter, 1987), a node u is influenced by each of its connections v by a weight w_{vu} where $\sum_{v\text{contact of } u} w_{vu} \leq 1$. Node u also has a threshold θ_u drawn randomly in [0,1]. At a given time period, u adopts the innovation if and only if $\theta_u \leq \sum_{v \text{ adapter and contact of } u} w_{vu}$. The intuition behind using innovation diffusion to model rumours is that a person may not believe in the rumour the first time they hear it, but they become more likely to believe in the rumour as they hear it from different people. In cascade models (Goldenberg, 2001), when a node adopts the innovation, each one of its contacts follows with some probability. Domingos and Richardson (Domingos and Richardson, 2001) suggested an optimisation problem of identifying the individuals that we can try to convince to adopt a new product or innovation in order to cause a cascade of adoption. This optimisation algorithm is NP-hard and Kempe et al. (2003) suggested an approximation algorithm to find these most influential nodes. Game-theoretic approaches (Morris, 2000) have also been used to study idea propagation in a social network. The intuition is that the utility that a person perceives from adopting an idea increases as more of the person's contacts adopt that idea. In the model, a person adopts at a given time period with a probability that increases with the number of contacts that have adopted it.

The literature contains empirical studies of spam; however, these studies are typically only marginally interested in spam diffusion in social networks. Kanich et al. (2008) find that spam campaigns are very profitable. Levchenko et al. (2011) characterise the infrastructure used to monetraize spam (Levchenko et al., 2011). Grier (2010) characterise spam on Twitter. For example, Grier et al. find that most spam on Twitter originates from hacked accounts.

Modelling and simulations haves been used to study the impact of cyber-attacks on various networks. Kundur et al. (2011) suggest an impact analysis framework for investigating the impact of attacks on the smart grid. There is considerable work on worm diffusion in computer networks (Zou et al., 2012; Wagner, 2003; Zou et al., 2003). Tang and Li (2011) analyse virus spread in wireless sensor networks using epidemic models. Karyotis (2006) introduce a probabilistic modelling framework for the diffusion of an energy-constrained mobile threat in a wireless ad hoc network and evaluate the impact of various parameters using simulations.

3 Background

3.1 Construct

Construct (Carley, 1991) is a powerful turn and agent-based simulation tool for investigating information diffusion and social change in complex socio-cultural systems. In agent-based simulation tools, the model is specified at the agent level, but the main interest is on emergent behaviour. Construct is turn-based in the sense that agents act in turn. Since Construct is a relatively complex tool, we only describe features relevant to this paper. In Construct, agents have information and are connected through a social network. Agents choose agents to interact with based on factors that include homophily and information seeking. Through these interactions, information is exchanged among agents. In homophily based interactions, human agents interact with human agents that have similar information. In information seeking interaction, human agents interact with human agents from whom they can learn the most information. Information is represented through an array, where each value is between 0 and 1. A value equal to 1 indicates that the agent has that information; a value equal to 0 indicates that the agent does not have this information whereas an intermediate value indicates that the agent has this information partially. For example, an information array (1, 1, 0.5, 0) indicates that the agent knows the first and second piece of information, knows half of the third piece of information, but does not know the fourth piece of information. At each time period, agents are associated in pairs, and each pair of agents interacts and exchanges information. More specifically, each agent has an initiation count that indicates the number of interactions the agent can initiate and a reception count that specify the number of interactions an agent can receive at a given time period. An agent can interact with himself in case no other agent is available for interaction. This self-interaction does not count towards the agent's reception count. Consider for example that agent A has initiation count equal to 3 and a reception count equal to 1, whereas agents B, C and D all have initiation and reception counts equal to 1. One possible pairing is $(A-B)$, $(A-C)$, $(B-D)$, $(C-A)$, $(A-A)$, $(D-D)$. A initiates the interaction with B, C and itself, B initiates the interaction with D, C initiates the interaction with A and D initiates the interaction with itself. Similarly, each agent receives an interaction initiated from a different agent once. During an interaction, each agent randomly chooses an information value from its information array and sends that value to its interaction partner. Upon receiving that value, this interaction partner updates its information array. Each agent places a transmission weight on each value in its information array. The information weight is the weight that the agent places on a given piece of information when choosing information to transmit. Consider for example that agent A places information weights (10, 1, 1, 0) on its information array (1, 1, 1, 1), then A has a probability $10/12$ of transmitting the first value, probability $1/12$ of transmitting the second value, probability $1/12$ of transmitting the third value and probability 0 of transmitting the fourth value. When A's interaction partner B receives an information value from A, B updates its information array. Consider for example that B has information array (0, 0, 0, 1) before interacting with A and that A decides to send the first value to B, then B's information array becomes (1, 0, 0, 1). Construct also allows specifying a learning rate and a forgetting rate per agent and per information. A learning rate is the probability that an agent that receives information learns that information. A forgetting rate is a value between 0 and 1 that specifies the rate at which an agent's information value decays over time. Assume that an agent has an information value v at a given time period and has forgetting rate fr, then the next time period the information value is $v*(1-fr)$. Consider for example that agent A has information array (1, 1, 1, 1) at time period t, and forgetting rate 0.1 for the first value and forgetting rate 0.6 for the other values. A's information array becomes (0.9, 0.4, 0.4, 0.4) at time $t+1$ and (0.81, 0.16, 0.16, 0.16) at time $t+2$. Besides information, agents also have transactive memory. We have not discussed transactive memory above in order to keep the discussion simple. Transactive memory represents agents' limited understanding of their environment. More specifically, an agent's transactive memory consists of what the agent thinks each of its contacts knows. What an agent thinks its contacts know does not necessarily match what these contacts know. This mismatch captures the bounded rationality notion. For example, agent A's transactive memory about agent B can be (1, 1, 0, 1), whereas agent B's information array is (1, 0, 0, 1). That is agent A thinks that agent B knows the first, second and fourth information, whereas in reality B only knows the first and fourth information. During an interaction, agents can also choose to transmit transactive memory. For example, agent A can send to agent B part of its transactive memory about agent C. Construct enables setting the probability that an agent chooses to transmit information and the probability that an agent chooses to transmit transactive memory.

3.2 Box–Behnken experiment design

Box–Behnken design (Box and Behnken, 1960) is one type of systematic experiment designs. A systematic experiment design helps choose the parameter combinations for which to run experiments in order to generate the response surface. Such design avoids the combinatorial complexity of running all possible combinations of the independent variable values. The Box–Behnken design is suited for quantitative variables and requires three values for each independent variable. This design suggests using parameter combinations that are the midpoints of the edges of the design space and the centre point. This design should be sufficient to fit a quadratic model.

4 Model

In this section, we describe the hacked account and human model and the human-only model. Hacked accounts initiate the rumour in the hacked account and human model, while humans initiate the rumour in the human-only model. The hacked account and human model has two types of agents: hacked agents correspond to people who have a hacked e-mail account and regular agents correspond to people who have no hacked e-mail account. Hacked agents initiate the rumour diffusion and are more aggressive about transmitting the rumour than regular agents. The human-only model also has two types of agents: initiator agents and regular agents. Initiator agents correspond to the people that initiate the rumour and regular agents correspond to the other people in the social network. Regular agents in the human-only model have the same behaviour as regular agents in the hacked account and human model. Initiator agents initiate the rumour propagation, but transmit the rumour at the same rate as regular agents.

4.1 Hacked account and human model

The model contains two types of agents: hacked agents that represent the people with hacked e-mail accounts and regular agents that represent people with no hacked e-mail account. Similarly, there are two types of information: regular information and the rumour. We use rumour as a generic term that can also designate spam or misinformation. We assume the absence of information conflicting to the rumour in the social network. This is a simplifying assumption that we intend to address in future work. The main difference between hacked agents and regular agents is that

- hacked agents initially know the rumour while the regular agents do not
- hacked agents are aggressive about transmitting the rumour.

In order to capture the fact that hacked agents send a very large number of e-mails, hacked agents have a higher initiation count than regular agents. Moreover, hacked agents place a transmission weight 1 on the rumour and transmission weight 0 on other information. This causes all e-mails originating from hacked e-mail accounts to contain the rumour. Regular agents, on the other hand, place equal transmission weight on the rumour and the regular information. Regular agents place a learning rate less or equal to 1, and a forgetting rate higher or equal to 0 on the rumour. This reflects that regular agents do not always believe the rumour and that regular agents lose interest in the rumour over time. On the other hand, hacked agents place a zero forgetting rate on the rumour. This causes the hacked agents to never lose interest in the rumour. As the hacked agents initially have the rumour and never lose interest in the rumour, the learning rate that hacked agents place on the rumour is irrelevant.

4.2 Human-only model

The human-only model has two types of agents: initiator agents and regular agents, and two types of information: the rumour and regular information. Regular agents in the human-only model have the exact same behaviour as regular agents in the hacked account and human model. Initiator agents have the same initiation count as regular agents. As a result, initiator agents transmit the same amount of messages as regular agents. Initiator agents have initially access to both the rumour and regular information. Initiator agents place a transmission weight 1 on the rumour and transmission weight 0 on the regular information. This causes the initiator agents to initiate the rumour transmission immediately at the start of the simulation. Finally, initiator agents place a non-zero forgetting rate on both regular information and the rumour.

5 Virtual experiment

In this section, we present our virtual experiment. Table 1 shows the experiment variables. The independent variables consist of the network size, network topology, network density, number of hacked (or initiator) agents, strategy for choosing hacked agents, learning rate and forgetting rate. We use a small, a medium and a large value for each of the quantitative variables, as required by the Box–Behnken experiment design. Such requirement does not apply to the categorical variables, namely the network topology and the strategy for choosing the hacked agents. The dependent variables consist of the number of agents that have the rumour over time and the maximum number of agents that have the rumour.

We first describe the independent variables. We use network sizes 100, 600 and 1100, which correspond to the size of a small, a medium and a large corporation respectively. Using million-node networks is computationally prohibitive within a complex simulation tool such as Construct. As future work, it would be interesting to investigate using such large networks within simpler simulation tools. We experiment with an Erdos-Renyi topology (Erdos and Renyi, 1960), a small-world topology (Watts and Strogatz, 1998) and a scale free topology (Barabasi and Albert, 1999). The Erdos-Renyi topology is a random topology almost never found in real social networks. We use the Erdos-Renyi topology as a baseline for comparison. The small-world topology is often used to model human social networks and the scale free topology is found in many online social networks (Mislove, 2007). In order to obtain consistent results across these topologies, networks are generated to have the same density. The density of a network is the ratio of links present out of the number of possible links in the network. We use values 2, 6 and 10%. The number of hacked agents represents the number of hacked e-mail accounts used to disseminate the rumour. We experiment with values 1, 3 and 5, which correspond to 1–5% in a network of 100 nodes, and to 0.09–0.45% in a network of 1100 nodes. These values are consistent with the size of high-profile password database breaches and account hacking. For example, 420,000 out of 28 million ($\approx$1%) Formspring passwords were leaked in 2012 (CNET, 2012) and 6.5 million out of about 150 million ($\approx$5%) LinkedIn passwords were leaked in 2010 (ZDNet, 2012). In early 2013, 250,000 Twitter accounts were hacked (abcNews, 2013). Next, we experiment with two strategies for choosing accounts to hack. In the random strategy, the hacked accounts are chosen randomly. This is, for example, the case in a password database breach or in a large scale non targeted phishing attack. The highest degree strategy corresponds to the case where the hacker targets the highest degree accounts using social-engineering and various attacks until the hacker is able to compromise these accounts. Hackers are interested in hacking and using high-degree accounts such as the fox news Twitter account (Guardian, 2011) because rumours from these accounts reach more people and have higher credibility. The learning rate is the probability that an agent adopts the rumour after hearing it from one of its contacts. In the experiment, we use values 20, 60 and 100%. Rumour theory identifies multiple factors that affect the likelihood that people spread

Table 1 Virtual experiment variables

Independent variables	*Number of test cases*	*Values used*
Network size	3	100, 600, 1100
Network topology	3	Small-world, scale free, Erdos-Renyi
Network density	3	2%, 6%, 10%
Number of hacked (or initiator) agents	3	1,3,5
Strategy for choosing hacked agents	2	Random, highest degree
Learning rate	3	20%, 60%, 100%
Forgetting rate	3	30%, 45%, 60%
Control variables		
Information array size	1	10
Probability of a regular agent to transmit information	1	0.8
Probability of a regular agent to transmit transactive memory	1	0.2
Homophily-based interaction	1	0.8
Expertise-seeking interaction	1	0.2
Time count	1	150
Dependent variables		*Variable range*
Ratio of agents that have the rumour over time	–	[0,1] each time period
Maximum ratio of agents that have the rumour	–	[0,1]

rumours. For example, rumours arise in ambiguous situations where people have a psychological need to understand (Fiske, 2004), but that reliable information is unavailable (Shibutani, 1966). Rosnow (1986) sent a questionnaire to faculty members of an American university asking them to list rumours they had heard recently and whether they had transmitted these rumours. The study occurred during major negotiations between the administration of the university and the faculty union. Rosnow et al. found that more credible rumours were more likely to be transmitted. They also found that 25.0% of low-credibility positive rumours and 31.4% of low-credibility negative rumours were transmitted, and that 71.4% of high-credibility positive rumours and 86.1% of high-credibility negative rumours were transmitted. Finally, the forgetting rate corresponds to the rate at which regular agents lose interest in the rumour and we use values 30%, 45% and 60%.

We now describe our control variables. The total number of information bits is 10 where 9 bits represent regular information and 1 bit represents the rumour. Initially, a given information value corresponding to regular information is set to 1 with probability 0.5 for all agents. On the other hand, the information value corresponding to the rumour is set to 1 for hacked agents and to 0 for regular agents. Next, hacked agents place a transmission weight of 1 on the rumour and a transmission weight of 0 on regular information, whereas regular agents place a transmission weight of 1 on each of their information bits. Furthermore, hacked agents have an initiation count equal to the network size, while regular agents have an initiation count of 1. The higher initiation count of hacked agents captures the large amount of e-mail transmitted by hacked e-mail accounts. Next, agents choose interaction partners based on knowledge similarity 80% of the time and in order to seek new knowledge 20% of the time. During interactions, agents exchange information 80% of the time and transactive memory 20% of the time. Finally, we run the experiment for 150 time periods. We determine the length of the experiment by running a few preliminary simulations and choosing the time length that enables observing the main behaviour.

6 Results

In this section, we discuss the results of our virtual experiment. More specifically, we compare the rumour diffusion dynamics in the hacked account and human model, and in the human-only model. Our results indicate that the rumour always reaches a higher number of agents in the hacked account and human model than in the human-only model. More surprisingly, the effect of some variables differs across the two models. For example, the network size has almost no effect in the hacked account and human model, but has considerable effect in the human-only model. Figure 1 shows the maximum ratio of agents that have the rumour in the two models for all parameter combinations of the Box–Behnken design. As can be seen, the maximum ratio of agents that have the rumour in the hacked account and human model is always higher than in the human-only model. This result is expected since hacked agents in the hacked account and human model aggressively disseminate the rumour. More specifically, in the hacked account and human model, hacked agents interact with up to all their contacts each time period and always choose to transmit the rumour. On the other hand, in the human-only model, at any given time period, each agent interacts with at most another agent. During these interactions, agents may or may not choose to transmit the rumour.

A surprising observation from Figure 1 is that the maximum ratio of agents that have the rumour behaves differently across different parameter combinations in the two models. For example, this maximum ratio oscillates in the human-only model, but remains almost constant in the hacked account and human model for parameter combinations 9 to 15. We see the same phenomenon for parameter combinations 21 to 24 and 27 to 36.

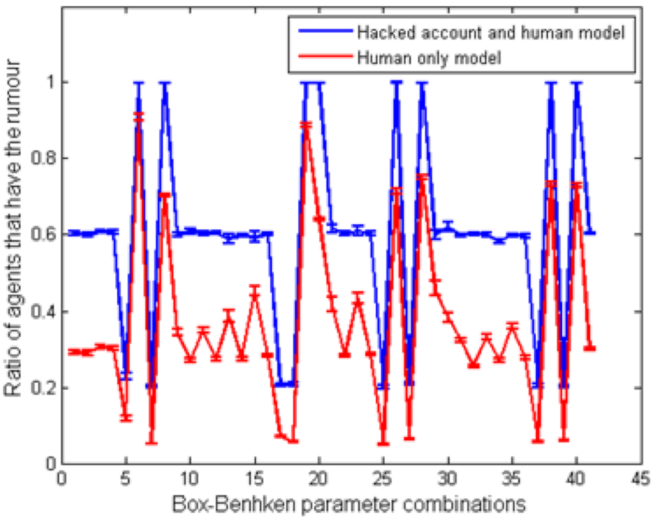
Table 2 presents the correlation between the maximum ratio of agents that have the rumour and the independent variables that are part of the Box–Behnken design. It can be seen that the learning rate has a strong effect on the results in the two models. The learning rate reflects the rumour credibility and agents' willingness to transmit the rumour. The rumour reaches a large fraction of the network when the learning rate is large, but remains localised when the learning rate is small. The network size and the forgetting rate impact the maximum ratio of agents that have the rumour in the human-only model, but not in the hacked account and human model. In order to gain more insight into these results, we first examine Figure 2 that shows the ratio of agents that have the rumour for all the parameter combinations of the Box–Behnken design. Subsequently, we examine Figures 3 and 4 that compare the effect of the network size and the forgetting rate on the diffusion in the two models.

Table 2 Correlation between the maximum ratio of agents that have the rumour and the quantitative independent variables

Independent variable	*Hacked account and human model*	*Human-only model*
Learning rate	0.995***	0.993***
Network size	–0.008	–0.185***
Forgetting rate	0.002	–0.117***
Number of hacked (initiator) agents	0.006	0.031
Network density	–0.001	0.004

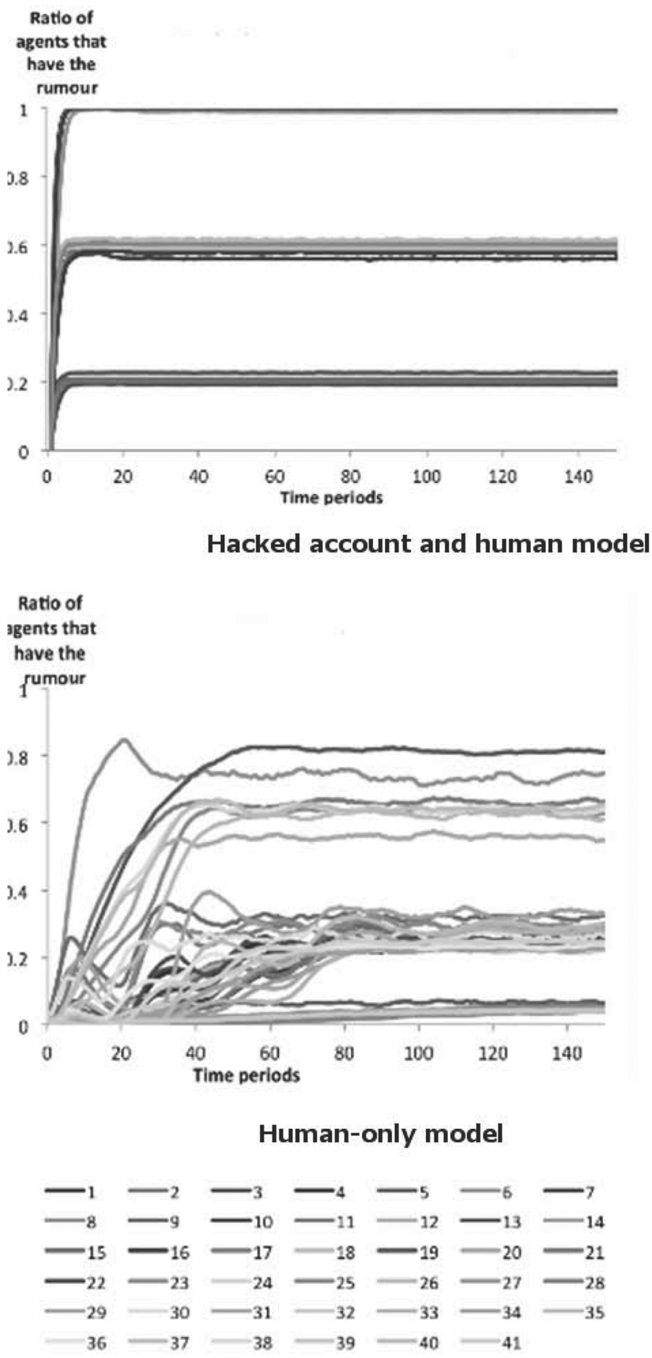
$^{***}p < 0.001$.

Figure 1 Comparison between the diffusion in the hacked account and human model, and the human-only model across all parameter combinations (see online version for colours)



From Figure 2, it can be seen that in the hacked account and human model, the ratio of agents that have the rumour increases very fast and reaches a maximum in a few steps. The ratio then remains stable at a maximum. On the other hand, in the human-only model, that ratio slowly increases and fluctuates before stabilising. In the hacked account and human model, the ratio of agents that have the rumour initially increases very fast because hacked agents send the rumour to a large number of agents. Later, the ratio of agents that have the rumour stabilises because, at a given time period, some agents loose interest in the rumour and the same number of other agents lose interest in it. In the human-only model, the ratio of agents that have the rumour fluctuates initially because the number of agents that learn the rumour and the number of agents that loose interest in the rumour evolve at different rates. More specifically, change in the number of agents that learn the rumour precedes in time change in the number of agents that loose interest in the rumour. As the two numbers start evolving at the same rate, the ratio of agents that have the rumour stabilises.

Figure 2 Behaviours of the ratio of agents that have the rumour over time for the different Box–Behnken parameter combinations



We now examine Figure 3 in order to gain more insight into the effect of network size. We see that the rumour reaches a higher ratio of agents for the small network size in the human-only model, but reaches the same ratio of agents in the hacked account and human model. As that ratio is equal to the number of agents that have the rumour divided by the network size, it is helpful to examine the number of agents that have the rumour over time. From Figure 3, we see that in the human-only model, the rumour initially reaches the same number of agents for the two network sizes. The number of agents that have the rumour is the same for the two network sizes because each initiator agent interacts with at most one other agent at any time period. Given that the ratio of agents is equal to the number of agents divided by the network size, the ratio of agents is larger for the smaller network size. From the figure, we also see that the number of agents that have the rumour stabilises later for the larger network. This is simply due to the fact that there are more agents in the larger network, and therefore the rumour can continue spreading. In the hacked account and human model, the rumour reaches the same ratio of agents. In this model, hacked accounts transmit the rumour to all their contacts. As the number of these contacts is proportional to the network size, the ratio of agents that have the rumour is the same for the two network sizes.

We now explain the difference in the effect of the forgetting rate in the two models. Figure 4 shows the ratio of the agents that have the rumour for forgetting rates 30% and 60%. In the human-only model, that ratio is smaller for the larger forgetting rate. When the forgetting rate is large, agents loose interest in the rumour faster and therefore transmit it to fewer agents. As a consequence, the rumour reaches less agents overall. In the hacked account and human model, the ratio of agents that have the rumour is unaffected by the forgetting rate. The rumour propagates very fast and reaches the maximum during the first few time steps before the forgetting rate has an effect.

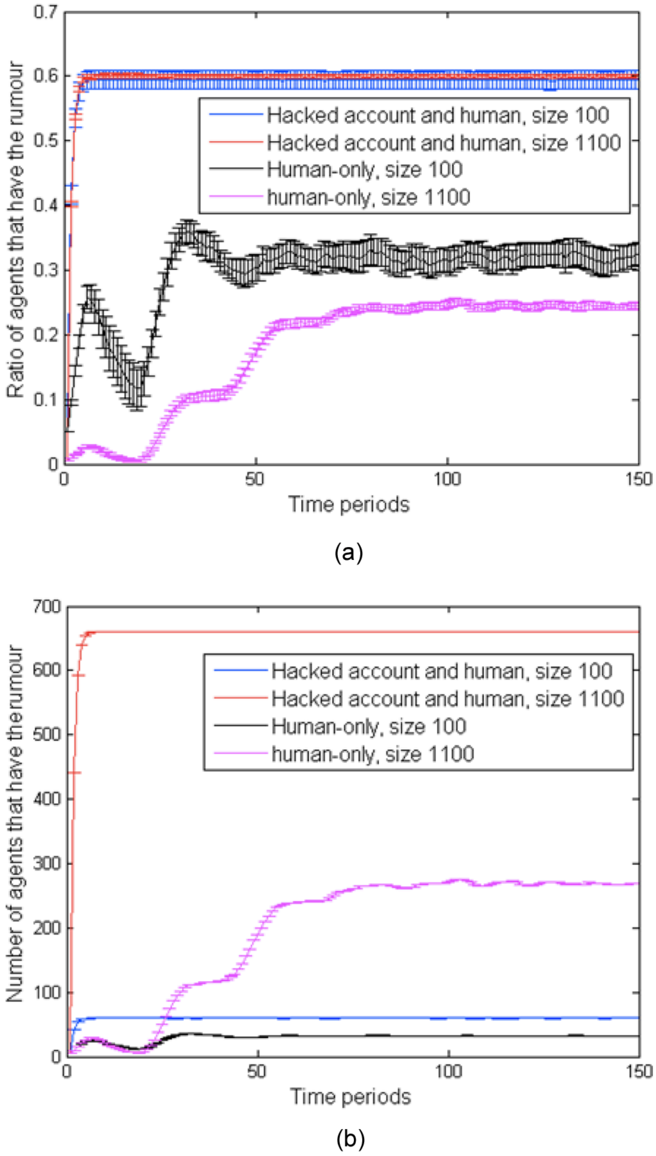
Figure 5 compares the effect of the random and highest degree strategies, and the effect of different network topologies. From the figure, we see that the strategy and the network topology do not significantly affect the rumour diffusion. These surprising results may be due to the short distance between nodes in the three types of networks.

6.1 Spam mitigation techniques

Multiple social networks and e-mail service providers implement spam mitigation techniques. We discuss how two major techniques, namely content filtering and account suspension affect the results of this paper.

The content filtering technique consists of examining e-mail content and flagging the e-mail as spam when the content is suspicious. The e-mail is then typically placed in a different folder and is less likely to be read by the recipient. Content filtering is used by most e-mail service providers, but not by Twitter (Grier, 2010). Capturing content filtering in our simulation can be accomplished by reducing the learning rate. Perfect content filtering is equivalent to a learning rate equal to 0. In this case, rumour diffusion does not occur in neither the hacked account and human model nor the human only model. Imperfect content filtering is equivalent to a smaller learning rate. In this case, the rumour remains localised and only reaches a small portion of the network. However, the rumour diffusion dynamics are still different across the two models as can be seen from Figure 2.

Figure 3 Comparison of the effect of the network size on the diffusion in the hacked account and human model, and the human model: (a) ratio of agents and (b) number of agents (see online version for colours)



Account suspension consists of blocking hacked accounts that send large amounts of e-mails. Account suspension does not affect human agents that spread the rumour as those agents send e-mails at a regular rate. In the hacked account and human model, the ratio of agents that have the rumour reaches the maximum within a few time periods. Thus, account suspension that is not extremely fast would not affect the maximum number of agents that receive the rumour. In case hacked accounts are suspended very early, the rumour diffusion will be slower, but will not stop as human agents will continue to spread the rumour.

Figure 4 Comparison of the effect of the forgetting rate on the diffusion in the two models. Small world network, random strategy, learning rate = 60%, number of hacked (initiator) agents, density = 6% (see online version for colours)

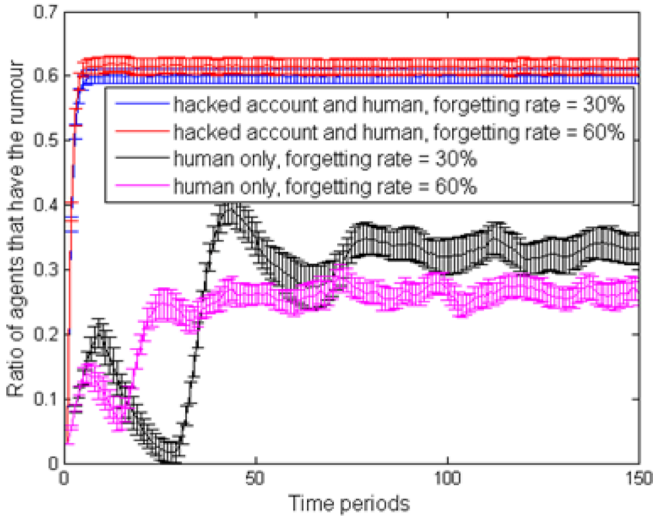
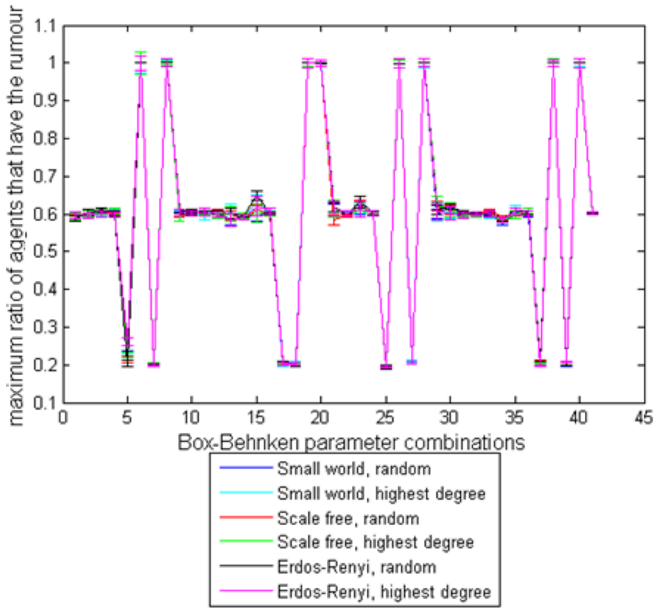


Figure 5 Effect of the network topology and the strategy for choosing hacked agents (see online version for colours)



7 Limitations and future work

In this paper, we focus on the case where hacked accounts initiate spam diffusion, and overlook the case where fake accounts initiate such diffusion. Spam transmitted by hacked accounts enjoys greater credibility than spam transmitted by fake accounts. Grier (2010) found that spammers mostly use hacked accounts in Twitter. As future work, it would be interesting to incorporate fake accounts in the simulation.

The model in this paper assumes that the more e-mails a hacked account sends to a person, the more likely is the person to think that the rumour is correct. This may be the case if the e-mails contain different 'facts' related to the rumour, but may not be necessarily the case if the content of the e-mails is exactly the same.

8 Conclusion

Spam diffusion in social networks is a major problem. In this paper, we investigate spam diffusion dynamics in social networks, where spam is initiated by hacked accounts. We build our spam diffusion model by modifying a standard diffusion model in order to capture the behaviour of hacked e-mail accounts. Our results show that when the behaviour of hacked accounts is captured, spam diffuses faster and reaches more people, and parameters like the network size and spam credibility affect the diffusion differently.

Acknowledgements

This work is supported in part by the Defense Threat Reduction Agency (DTRA) under the grant number HDTRA11010102, and tthe Army Research Office (ARO) under grants W911NF1310154 and W911NF0910273, and the centre for Computational Analysis of Social and Organisational Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of DTRA, ARO or the US government.

The authors would like to thank Geoffrey Morgan, Michael Lanham and Kenneth Joseph for useful feedback and interesting discussions.

References

abcNews (2013) *250,000 Twitter Accounts Hacked: Don't Panic, Here's What To Do*, http://abcnews.go.com/blogs/technology/2013/02/250000-twitter-accounts-hacked-dont-panic-heres-what-to-do/, Last accessed: October, 2013.

Barabasi, A.L. and Albert, R. (1999) 'Emergence of scaling in random networks', *Science*, Vol. 286, pp.509–512.

Box, G.E.P. and Behnken, D.W. (1960) 'Some new three level designs for the study of quantitative variables', *Technometrics*, Vol. 2, pp.455–475.

Carley, K.M. (1991) 'A theory of group stability', *American Sociological Review*, Vol. 56, pp.331–354.

CNET (2012) *Formspring Disables User Passwords in Security Breach*, http://news.cnet.com/8301-1009_3-57469944-83/formspring-disables-user-passwords-in-security-breach/, Last accessed: October 2013.

ConsumerReports.org (2011) *Report: 600,000 Facebook Log-Ins Compromised per Day*, http://www.consumerreports.org/cro/news/2011/10/report-600-000-facebook-log-ins-compromised-per-day/index.htm, Last accessed: October, 2013.

Domingos, P. and Richardson, M. (2001) 'Mining the network value of customers', *Seventh International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, August, pp.57–66.

Erdos, P. and Renyi, A. (1960) *On the Evolution of Random Graphs*, Publications of the Mathematical Institute of the Hungarian Academy of Sciences, p.5.

Fiske, S.T. (2004) *Social Beings: Core Motives in Social Psychology*, John Wiley and Sons, New York.

Go Hacking (2008) http://www.gohacking.com/hack-email-account-password/, Last accessed: October 2013.

Goldenberg, J., Libai, B. and Muller, E. (2001) 'Talk of the network: a complex systems look at the underlying process of word-of-mouth', *Marketing Letters*, Vol. 12, pp.211–223.

Granovetter, M. (1987) 'Threshold models of collective behavior', *American Journal of Sociology*, Vol. 83, pp.1420–1443.

Grier, C., Thomas, K., Paxson, V. and Zhang, M. (2010) '@spam: the underground on 140 characters or less', *Conference on Computer and Communications Security (CCS)*, October, Chicago, IL.

Guardian (2011) *Fox News's Hacked Twitter Feed Declares Obama Dead*, http://www.theguardian.com/news/blog/2011/jul/04/fox-news-hacked-twitter-obama-dead, Last accessed: October, 2013.

Hacker The Dude (2013) *How to Hack any Email Accoount*, http://hackerthedude.blogspot.com/2009/06/how-to-hack-any-email-account.html, Last accessed: October 2013.

Kanich, C., Kreibich, C., Levchenko, K., Enright, B., Voelker, G.M., Paxson, V. and Savage, S. (2008) 'Spamalytics: an empirical analysis of spam marketing conversion', *Conference on Computer and Communications Security (CCS)*, October, Alexandria, VA.

Karyotis, V., Papavasiliou, S., Grammatikou, M. and Maglaris, V. (2006) 'A novel framework for mobile attack strategy modelling and vulnerability analysis in wireless ad hoc networks', *International Journal of Security and Networks*, Vol. 1, pp.255–265.

Kempe, D., Kleinberg, J. and Tardos, E. (2003) 'Maximizing the spread of influence through a socia network', *Ninth International Conference on Knowledge Discovery and Data Mining*, ACM SIGKDD, August, Washington DC, pp.137–146.

Kundur, D., Feng, X., Mashayekh, S., Liu, S., Zourntas, T. and Butler-Purry, K.L. (2011) 'Towards modeling the impact of cyber attacks on a smart grid', *International Journal of Security and Networks*, Vol. 6, pp.2–13.

Levchenko, K., Pitsillidis, A., Chachra, N., Enright, B., Felegyhazi, M., Grier, C., Halvorson, T., Kanich, C., Kreibich, C., Liu, H., McCoy, D., Weaver, N., Paxson, V., Voelker, G.M. and Savage, S. (2011) 'Click trajectories: End-to-end analysis of the spam value chain', *IEEE Symposium on Security and Privacy*, May, Oakland, CA.

Metaxas, P.T. and Mustafaraj, E. (2012) 'From osbcurity to prominece in minutes: political speech and real-time searches', *Web Science Conference*, Evanston, IL.

Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P. and Bhattacharjee, B. (2007) 'Measurement and analysis of online social networks', *Internet Measurement Conference (IMC)*, San-Diego, October.

Morris, S. (2000) 'Contagion', *Review of Economic Studies*, Vol. 67, pp.57–78.

Ratkiewicz, J., Conover, M.D., Meiss, M., Goncalves, B., Flammini, A. and Menczer, F. () 'Detecting and tracking political abuse in social media', *Fifth International AAAI Conference on Weblogs and Social Media*, July, Barcelona, Spain, pp.297–304.

Rosnow, R.L., Yost, J.H. and Esposito, J. (1986) 'Belief in rumor and likelihood of rumor transmission', *Language and Communication*, Vol. 6, pp.189–194.

Shibutani, T. (1966) *Improvised News: A Sociological Study of Rumor*, Bobbs-Merrill, Indianapolis, IN.

Tang, S. and Li, W. (2011) 'An epidemic model with adapative virus spread control for wireless sensor networks', *International Journal of Security and Networks*, Vol. 6, pp.201–210.

Valente, T. (1996) 'Network models of the diffusion of innovations', *Computation and Mathematical Organization Theory*, Vol. 2, pp.163, 164.

Wagner, A., Dubendorfer, T., Plattner, B. and Hiestand, R. (2003) 'Experiences with worm propagation simulations', *Proceedings of the ACM Workshop on Rapid Malcode (WORM)*, San Diego, CA, pp.34–41.

Watts, D. and Strogatz, S. (1998) 'Collective dynamics of 'small-world' networks', *Nature*, Vol. 393, pp.440–442.

Wu, F., Huberman, B.A., Adamic, L. and Tyler, J. (2003) 'Information flow in social groups', *Physica A: Statistical Mechanics and its Applications*, Vol. 337, pp.327–335.

Yahoo! News (2010) *Fake Tsunami Warning Sent from Hacked Twitter Account*, http://news.yahoo.com/fake-tsunami-warning-sent-hacked-twitter-account.html, Last accessed: October, 2013.

Zanette, D. (2002) 'Dynamics of rumor propagation on small-world networks', *Physical Review E*, Vol. 65, pp.041908 1-9

ZDNet (2012) *LinkedIn password Breach: How to Tell if You're Affected*, http://www.zdnet.com/blog/btl/linkedin-password-breach-how-to-tell-if-youre-affected/79412, Last accessed: October, 2013.

Zou, C., Gong, W. and Towley, D. (2002) 'Code red worm propagation modeling and analysis', *Proceedings of the ACM Workshop on Rapid Malcode (WORM)*, Washington, DC, pp.138–147.

Zou, C., Gong, W. and Towsley, D. (2003) 'Worm propagation modeling and analysis under dynamic quarantine defense', *Proceedings of the ACM workshop on Rapid Malcode (WORM)*, San Diego, pp.51–60.

Appendix

Table 3 contains parameters used in our virtual experiment.

Table 3 Experiment design based on the Box–Behnken experiment Design

	Number of compromised agents	*Network density*	*Network size*	*Learning rate*	*Forgetting rate*	*Number of replications*
1	1	2	600	60	45	50
2	1	10	600	60	45	50
3	5	2	600	60	45	50
4	5	10	600	60	45	50
5	3	6	100	20	45	50
6	3	6	100	100	45	50
7	3	6	1100	20	45	50
8	3	6	1100	100	45	50
9	3	2	600	20	30	50
10	3	2	600	100	60	50
11	3	10	600	60	30	50
12	3	10	600	60	60	50
13	1	6	100	60	45	50
14	1	6	1100	60	45	50
15	5	6	100	60	45	50
16	5	6	1100	60	45	50
17	3	6	600	20	30	50
18	3	6	600	20	60	50
19	3	6	600	100	30	50
20	3	6	600	100	60	50
21	3	2	100	60	45	50
22	3	2	1100	60	45	50
23	3	10	100	60	45	50
24	3	10	1100	60	45	50
25	1	6	600	20	45	50
26	1	6	600	100	45	50
27	5	6	600	20	45	50
28	5	6	600	100	45	50
29	3	6	100	60	30	50
30	3	6	100	60	60	50
31	3	6	1100	60	30	50
32	3	6	1100	60	60	50
33	1	6	600	60	30	50
34	1	6	600	60	60	50
35	5	6	600	60	30	50
36	5	6	600	60	60	50
37	3	2	600	20	45	50
38	3	2	600	100	45	50
39	3	10	600	20	45	50
40	3	10	600	100	45	50
41	3	6	600	60	45	300