An Experimental Framework for Characterizing Adversarial Behavior in Social Networks

Pivithuru Wijegunawardana[1], Mario Taglic[1], Ralucca Gera[2], and Sucheta Soundarajan[1]

[1] Syracuse University, Department of Electrical Engineering & Computer Science, {ppwijegu, mtaglic, susounda}@syr.edu
[2] Naval Postgraduate School, Department of Applied Mathematics, RGera@nps.edu

Abstract.

We introduce the Adversarial Social Network Analysis game (ASNA game), an experimental framework to study the adversarial behavior of covert groups in social networks. We conduct experiments on Amazon Mechanical Turk using the proposed framework, which models interactions between a data collector and members of an adversarial group as a network game. Participants of the experiments play as adversarial nodes in the game. Our goal is to understand the factors that motivate an individual to report network or attribute data dishonestly. We demonstrate the validity of the ASNA game framework by showing that participants understand and pay attention to key elements of the game, they try to maximize their objective in the game, and they show loyalty to the adversarial group.

Keywords: Adversarial behavior · network games · behavior experiments.

1 Introduction

Social network analysis is often used to gain fundamental insights into human social interactions. For example, one can identify the most influential individuals in a network, understand how information spreads through groups, and characterize the community structure of a population. Typically, when one performs such analysis, one assumes the availability of accurate information about the network structure and people who are part of it.

However, in certain applications, such assumptions about data accuracy may not hold. In particular, we consider the setting where there is a covert group (such as a criminal gang or terrorist cell) hiding among civilians in a social network. The nodes representing these people are adversarial against those who seek to collect or analyze their data. Thus, the covert nodes may misreport data about their network structure or attributes (their own or others') to a data collector in order to hinder such analysis and hide the identities of themselves and their compatriots. While most data analysis tasks must deal with noisy data, this type of adversarial behavior seeks to deliberately deceive an analyst, and may present different challenges than simply random noise.

To better understand adversarial network structure, one should be able to characterize *how* individuals in these networks may attempt to deceive a data colletor. Such information can be used to design network analysis algorithms that are more robust against purposeful misinformation, characterize the network structures that promote adversarial behaviors, identify regions of the network for which more accurate data should be collected (by, e.g., recruiting informants from the group), predict the evolution of criminal groups, and so on.

In this work, we propose the Adversarial Social Network Analysis Game (`ASNA game`) framework to study the deceptive behaviors of adversarial nodes in social networks. We formulate the framework as a network game, where a data collector is attempting to reveal true information about nodes in the network while individuals in an adversarial group are attempting to disrupt the process. The `ASNA game` framework can be used to experimentally examine a wide variety of questions, including: 1) To what extent, and how, does the network structure influence the deceptive behavior of adversarial nodes?, 2) What level of incentive should a data collector provide to an adversarial node in order to obtain more useful, accurate data? 3) Does loyalty of individuals to the adversarial group or respect for the hierarchy of the organization play a role in deceptive behavior?

We use the `ASNA game` framework to conduct behavioral experiments using participants recruited from Amazon Mechanical Turk (mTurk). These participants play as members of an adversarial group. Our ultimate goal is to use `ASNA game` framework to propose a model that can predict a node's reporting behavior for a specific network setting.

In this work, we introduce the `ASNA game` framework and then discuss its implementation on mTurk. We then confirm the validity of `ASNA game` framework to study adversarial behavior of people by showing 1) Participants understand their role and aspects of the framework, 2) Participants understand their objective in a game and, 3) Participants act as members of the adversarial group.

2 Literature Review

Many studies on adversarial behavior come from the criminology, psychology, and sociology literature, and are conducted on criminals under interrogation. Some studies contrast strategies used by guilty suspects vs innocent suspects in a police interrogation [4], identify the stage at which interrogation evidence should be displayed to detect deception [5], and compare lie-telling strategies by people with criminal experience [12]. Although such studies are plentiful, they generally do not consider network structure. However, as shown by Galeotti [3] and Wong et al. [15], criminals usually form networks to be able to operate and adjust their networks to hide information.

Recent research on covert network analysis has looked specifically at applying social network analysis to analyze dark networks [2,6,7]. Others have considered using social network analysis techniques to disrupt criminal networks [10, 11]. These works mainly consider standalone criminal networks which are not embedded into regular social network, and work under the assumption of correct information. Wijegunawardna et al. [14] propose algorithms to identify criminals in a dark network under the deceptive behavior of individuals. Even though they consider possible inaccuracies in data, they use synthetic models to simulate these inaccuracies. The goal of the present work is to provide such algorithms realistic models of adversarial behavior.

We conduct our experiments using participants hired from Amazon Mechanical Turk (AMT). Recently, AMT has become popular among researchers running human experiments due to the low cost of and easy access to large subject pool [8]. Paolacci et al. [9] show that results obtained from running experiments on AMT are comparable to that of running lab experiments. However, AMT experiments can suffer from bots acting to be human subjects. Mason et al. [8]

suggest using filtering checks to ensure that mTurkers understand and pay attention to experiment details. There have not been many works on conducting network experiments on AMT. Suri et al. [13] conducted a networked public goods game experiment on AMT to evaluate the effect of network structure in human cooperation. Our work focuses on studying adversarial behavior of nodes in a network using AMT participants.

3 Adversarial Social Network Analysis Game

We model the `ASNA game` as a variation of the popular Werewolf/Mafia party games [1], in which a minority of players collude against the majority. In the `ASNA game`, a pack of werewolves (`Red` nodes), representing members of an adversarial group, is hiding among innocent civilian villagers (`Blue` nodes) in a village social network, as shown in Figure 1. (For ease of reading, in this paper, we use `Black` nodes to represent `Red` nodes and `White` nodes to represent `Blue` nodes in the figures. In our actual experiments, their true colors are red and blue.) A sheriff (data collector) begins an investigation to identify the werewolves, who attempt to evade detection. In the current implementation, participants play as werewolves only.

Overview of Gameplay: Currently, the game is a single player game, though future versions will allow for multiple players. Each participant (mTurkers) plays as a "YOU" node, a member of the adversarial group, as shown in Figure 1. The "You" node has just been investigated by the data collector, and the participant must thus decide whether to report each of her uninvestigated neighbors as `Red` or `Blue` nodes. Each choice is associated with potential rewards and penalties. She may choose to protect some members of the werewolf pack by claiming that they are villagers, and in doing so receive a reward from the pack; but if the data analyst subsequently investigates that neighbor and realizes her lie, she will be penalized. In contrast, if she betrays another werewolf, the pack will penalize her but the data analyst may reward her. To maximize her overall payoff, she must attempt to predict whether a node will subsequently be investigated by the data analyst.

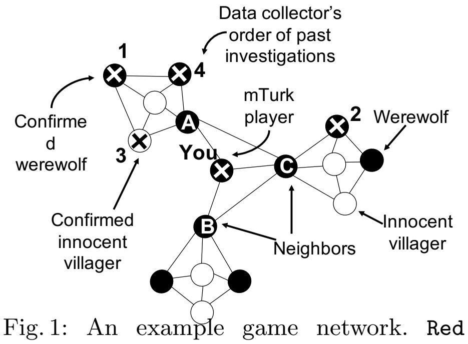
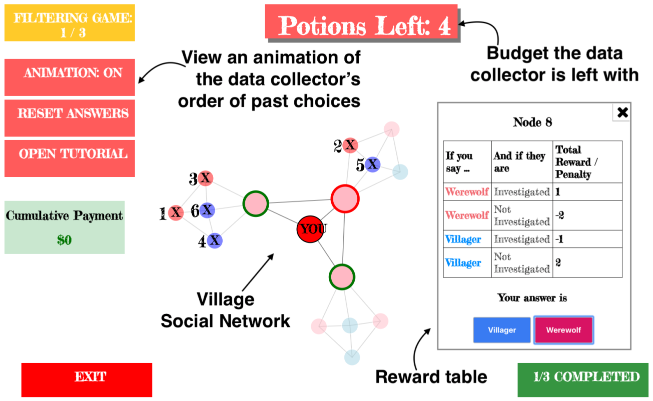


Fig. 1: An example game network. `Red` nodes are werewolves and `Blue` nodes are villagers. Dark colored nodes have already been investigated. The numbers by the investigated nodes show the order of investigator past choices. mTurk participant plays as the "YOU" node in the network. Participants need to report colors of the green circled neighbors of the "YOU" node in the network.

UI: In this version of the game, the player sees the full network structure, including the village's social network, the order of data collector's past choices, shown as numbers beside nodes that are already investigated, and the budget the investigator is left with, which we refer to as potions, in the game UI, as shown in Figure 2. At the beginning of each game, players see an animation showing which other individuals have been investigated before the player. This helps the player to determine the data collector's strategy.

Using this information, the player must decide whether to report each neighbor as a werewolf or a villager. Players report each neighbor as a `Red` or `Blue` node by clicking the Werewolf or Villager button in the rewards table.



If you say ...	And if they are	Total Reward / Penalty
Werewolf	Investigated	1
Werewolf	Not Investigated	-2
Villager	Investigated	-1
Villager	Not Investigated	2

Fig. 2: User interface the participants would interact with to report their answers about neighbors. The rewards table opens when a participant clicks on a neighbor and shows possible rewards she can earn for each possible answer. Potions left indicates how many other nodes, the data collector can investigate after the participant.

Rewards and Penalties: To see the rewards and penalties associated with the possible responses about a neighbor, the player clicks on that node. This action shows a matrix similar to that shown in Table 1. Rewards and penalties are dependent on 1) the true colors of the neighbors, 2) Whether a participant decide to report the truth or lie about neighbor's color and, 3) whether the data collector ultimately investigates the neighbor. This last factor is not known to the player- he or she must attempt to predict it.

Participants earn a reward from the werewolf pack for reporting false information about a `Red` node- i.e., if they lie to protect a fellow werewolf. Similarly, they get a penalty if they report the truth- i.e., betray a fellow werewolf. On the other hand, participants earn a reward from the data collector if they report the true colors of a werewolf, but get a penalty for lying. Importantly, the data collector awards rewards and penalties only if he subsequently investigates neighbors and confirm whether the participant has lied or told the truth. Thus, the player must attempt to predict the data collector's future actions when deciding her responses. The total reward a participant earns in a game converts to the bonus payment she gets for participating in the experiment. The specific values of rewards and penalties vary depending on the conditions being tested, and correspond to total payouts of roughly \$0.10 - \$0.70.

Data Collector: In our current experiments, we assume that the goal of the data collector is to investigate, and thus identify, as many Red nodes as possible. The data collector follows some strategy when selecting individuals to investigate (e.g., selecting those who are connected to to the most known adversarial nodes,

		Participant			
		Red Neighbor		**Blue Neighbor**	
		Truth	**Lie**	**Truth**	**Lie**
Investigator	**Investigates**	+ reward	- penalty	+ reward	- penalty
	Doesn't investigate	0			
Adversarial Group		- penalty	+ reward	0	

Table 1: Possible rewards and penalties participants can earn from the data collector and the adversarial group.

randomly selecting nodes, etc.). The data collector has some limit b on the number of nodes that she can investigate. Following the Werewolf game, we refer to each investigation as a "potion". As described above, to incentivize correct responses, the data collector provides penalties or rewards in response to false or true query responses. In our experiments, the data collector is controlled by a bot, and varying the strategy and incentives is part of our experimental conditions that we control.

Goals: The player's task is to decide what to report about each neighbor's color. Presumably, players will generally be motivated by the goal of maximizing their payoff, which corresponds to actual money. However, in some cases, players may show behaviors that are not optimal for this goal. For example, participants may decide to lie about civilian neighbors in order to mislead the investigation and protect more red neighbors, even if this does not lead to the greatest payoff.

In the tutorial preceding game play, to help participants make decisions, they see the following guidelines: 1) They should consider the number of potions left, because this plays a role in the likelihood that the data collector will investigate a particular neighbor, 2) They should attempt to identify any pattern in the data collector's past strategy, because this can help predict which nodes the data collector will investigate next, and 3) They should evaluate the rewards and penalties associated with each choice. The players need to pay attention to how likely it is that she will get caught lying to the investigator when deciding what to report.

Current Rules: The framework allows for many different rule sets. In the current version, the data collector iteratively selects individuals from the network (in accordance with some query strategy) for investigation, and (1) In each query, the data collector determines whether the node is `Red` or `Blue` with perfect accuracy, (2) When the data collector queries the selected node as to the colors of its neighbors (i.e., asks the individual under investigation whether her friends are werewolves or civilians), the queried node may lie.

MTurk Implementation: We implement the `ASNA game` on Amazon Mechanical Turk. In our current experiments, only one member of the adversarial group is played by a human mTurker; the other players are controlled by bots. These bots are placeholders for rest of the nodes in the village social network and do not add any information to the game play in the current version of the game. mTurkers first participate in a tutorial which explains the game interface, gameplay, the stakeholders in the game, and the player's role, as well as giving example games. After completing the tutorial, players take part in three filtering games, and must pass at least two in order to continue to validate that they understand their role.

4 Experiments and Results

In this section, we demonstrate the validity of the `ASNA game` by showing that players 1) Pay attention to key elements of the `ASNA game`, 2) Understand the objectives of the game and, 3) Show loyalty to the `Red` group. We set up the game so that at least 20 people participated as the "YOU" (`Red`) node in each experiment. Participants are able to take part in any number of experiments after completing the tutorial and passing the filtering exams.

4.1 Participants pay attention to key elements of the ASNA game

There are three key elements to the `ASNA game` that participants need to pay attention to in order for results to be valid: 1) The data collector's past choices 2) How many other individuals the data collector can investigate after the current participant and, 3) The rewards and penalties that the participant would get for each answer choice. We use three filtering games to ensure that participants understand and pay attention to these factors. In the first filtering game, we test if participants pay attention to whether there is a pattern to data collector choices. We use a path network with the data collector investigating one node at the time, in the order listed on the path. Given this setting, the data collector's next choice is likely to be the participant's neighbor. If the participants are paying attention to the details of the game, they should always report the true color of the neighbor.

	Passed	**Passed only one**	**Passed only two**	**Passed all three**
Filtering 1: Data collector choices The player must accurately predict that the data collector will investigate his/her neighbor next based on data collector's pattern of choices	63%	25%	36%	29%
Filtering 2: Budget left Enough budget to investigate all nodes in the network	55%			
Filtering 3: Reward table Rewards and penalties are such that the player will get a higher reward by reporting the true colors.	69%			

Table 2: Percentage of participants who passed each filtering game.

In the second filtering game, the data collector has enough query budget to investigate every node in the network. Therefore, participants should always report correct colors of the neighbors.

In the third filtering game, we check whether participants consider rewards and penalties when providing their answers. We allocate rewards and penalties to neighbors such that they would always get a higher reward by reporting true colors regardless of whether the data collector investigates them or not.

Table 2 shows statistics about how many participants passed each filtering game. If participants were to guess answers fully at random, there would be a 50% chance, 25% chance and 25% of passing filtering game 1, 2, and 3, respectively, and the probability of passing at least two filtering exams is 0.25. However, 211 mTurkers completed all three filtering games, and 139 (66%) passed at least two filtering games, and moved on to the next stage of the experiment. These filtering games help ensure that participants understand the game.

4.2 Participants understand objectives of the game

To further assess whether participants understand the game, we performed three experiments to evaluate whether the participants try to maximize their reward in the game. We designed experiments corresponding to this hypotheses, as follows:

Hypothesis 1: *People are more likely to report false information if they think that the data collector will not discover the truth.*

The `ASNA game` framework can convey information about the risk of the data collector uncovering the truth through the animation showing the data collector's past investigations. For example, if the data collector past choices seem to be random, she poses a little threat to the participant as compared to the case where the data collector is making choices near participant's neighbors. We use three different settings to test this hypothesis.

Setting 1: The data collector follows a clear pattern of investigation vs. a random selection of nodes

If the number of potions remaining is small compared to the number of uninvestigated nodes, we expect to see people misreport colors of their neighbors more often when the data collector has made random choices in the past as compared to when the data collector has followed a clear pattern of investigation that is likely to lead to the participant's neighbors. We use two networks to test this hypothesis. In one network, the data collector investigates nodes selected at random, and in the other network, the data collector investigates nodes row by row in the network. Rewards and penalties are the same for respective nodes in both networks. We randomly assign participants to one of the two networks to evaluate whether there is a difference in likelihood of misreporting colors between the two networks.

Participants reported false information about `Red` nodes twice as often in the random choice network as compared to when the data collector is following the row-by-row pattern. We confirm using a one tailed t-test at $p = 0.05$ significance that the participants on average reported false information about `Red` nodes in the random choice network significantly more compared to when the data collector follows some pattern.

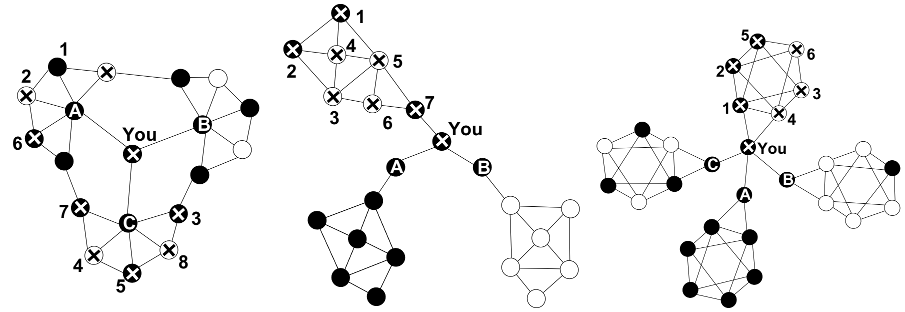
Setting 2: Data collector investigates some parts of the network more compared to others

In this setting, we test Hypothesis 1 by using the network shown in Figure 3a. The data collector has investigated each neighbor's neighborhood to a varying degree. We expect participants to identify those neighbors whose neighborhoods have been investigated less thoroughly, and thus provide false information about those neighbors more often. We keep rewards and penalties the same for all three neighbors.

Participants reported that C is not a `Red` node more often compared to A and B, even though all of them are indeed `Red` nodes. The only difference between these three neighbors is in the fraction of their neighborhoods that have been investigated. Using a t-test we confirm that likelihood of misreporting B is significantly lager than the likelihood of misreporting A and C at the $p = 0.05$ significance level.

Setting 3: All neighbors have similar anticipated rewards

We created another network to evaluate how participants balance between risk



(a) Percentage of neighbors investigated (b) Bridge between Red and Blue groups (c) Protecting neighbors leading to more red nodes

	# participants	A	B	C	A, B	B, C	A, C	A, B, C
% of nbrs investigated	24	54%*,†	67%*,‡	38%†,‡	46%	29%	33%	29%
Bridge between `Red` and `Blue` groups	20	50%*	15%*		5%			
Protect nbrs leading to more `Red` nodes	36	39%†	42%‡	31%†,‡	31%	25%	25%	31%

Fig. 3 & Table 3: Network a) tests whether the likelihood of adversarial nodes misreporting data changes (based on how participants perceive the likelihood of the data collector discovering the truth) and, networks b) and c) test whether participants show loyalty to the `Red` group. Table shows the percentage of participants, who have misreported information about their neighbors in the networks shown. Statistically significant differences from a one tailed t-test at 0.05 significance are marked with $*, \dagger, \ddagger$ symbols. Cells with the same symbol represent that the larger value is significantly larger than the smaller value.

and reward. In this network, participant has there neighbors A, B, and C who are at high, medium, and low risk of being investigated, respectively. Moreover, they have rewards and penalties of $\pm 5, \pm 3$, and ± 1, respectively. In other words, the higher the risk, the better the potential reward and the worse the potential penalty. The ultimate payoffs would be similar for all three neighbors when we consider the risk vs reward. Therefore, if participants are trying to maximize the reward they would earn, they should lie equally about all three neighbors. Experiments conducted using this network confirm this hypothesis since participants have not lied about any of the three neighbors significantly more than any other neighbor.

These three experiments confirm our hypothesis that participants understand their objectives in the game. Therefore, the experiments validate that the `ASNA game` framework can be used to understand deceptive behavior of adversarial nodes.

4.3 Participants are loyal to the Red group

Even though we want to study behavior of adversarial nodes, our experiment subjects consists of workers from AMT, who may not be people with an inherent adversarial mindset. However, even though the participants may not themselves be criminals, the framework allows us to study aspects of general human behavior. One trait of interest is that of members of some group is being loyal to that group.

Hypothesis 2: *If nodes show loyalty to their group, they would take risks and report false information about neighbors to protect members of their own group*

We test this hypothesis in two different settings. In the first setting, we position the player as a bridge between a `Red` group and a `Blue` group, to see if she tries to direct the data collector towards the `Blue` group and protect the `Red` group. In the second setting, we test whether the size of the `Red` group matters to the participant in the decision process, with players preferring to protect larger `Red` groups.

Setting 1: Bridge node between neighbors leading to a `Red` group and a `Blue` group
Figure 3b shows the experiment we designed to test Hypothesis 2. A and B both have same rewards and penalties, and the investigator order of past choices doesn't provide any useful information about which of them would be investigated next. Participants reported false information about A three times more often. According to the network structure, reporting the truth about A would lead the investigation towards a large `Red` compared to B. A t-test with $p = 0.05$ significance level confirms that participants are likely to provide false information about A significantly more often compared to B. This shows that participants are loyal to the `Red` group even if this does not affect their reward.

Setting 2: Protect larger `Red` group
We formulate the network in Figure 3c. The "YOU" node is adjacent to nodes A, B, and C, which lead to six, three, and one `Red` nodes, respectively. Rewards and penalties are the same for all neighbors. Table 3 shows that participants lied about nodes A and B significantly more compared to node C (significant at $p = 0.05$). However, we do not see a significant difference between the likelihood of lying about A, compared to B. This may be due to participants perceiving that both A and B would lead to similar `Red` groups. Since participants have lied significantly more about A and B compared to C, we can still conclude that participants try to protect neighbors that lead to a larger `Red` group. However, we cannot quantify the how large the `Red` group should be to observe such behavior.

These two experiments show that participants try to protect the `Red` groups even if that doesn't necessarily maximize their objective. We can conclude that participants are loyal to the `Red` group and try to protect the members since they act as members of the group.

5 Discussion and Conclusion

We propose the `ASNA game` framework to study adversarial behavior of nodes in a social network. The `ASNA game` framework is modeled as a network game played between a data collector and members of an adversarial group. By varying aspects of the game, we evaluate how the network structure, rewards and penalties, and data collection behavior influence adversarial behavior. Initial analysis using Amazon Mechanical Turk shows that 1) Participants understand their role in the game, and 2) Participants show loyalty to the group. Findings from this type of analysis may be helpful in designing network analysis algorithms that are robust to targeted misinformation, or in understanding the behavior of covert groups in general.

One drawback of our current work is that we recruit general workers from AMT to participate, rather than criminals. Thus, their mindset is different, creating different behaviors. However, we show that even with these workers, participants are loyal to the `Red` group and provide misinformation to the data collector to protect `Red` nodes showing that they in fact act adversarial towards the data collector without any necessary gain in rewards.

We intend to extend the present work by exploring other factors, such as the stated hierarchical ranks of other nodes, the existence of herding behavior, tit-for-tat behavior, more variations on network structure, how rewards and penalties can affect node behavior and so on. Additionally, we will extend the current work to a multi-player setting, allowing us to better understand group dynamics.

6 Acknowledgements

This material is based upon work supported by the U.S. Army Research Office under grant number # W911NF1810047. We are grateful for assistance and advice provided by Kiran Lakkaraju, Ricky Laishram and Katchaguy Areekijseree.

References

1. Braverman, M., Etesami, O., Mossel, E., et al.: Mafia: A theoretical study of players and coalitions in a partial information environment. The Annals of Applied Probability **18**(3), 825–846 (2008)
2. Chen, H., Chung, W., Xu, J.J., Wang, G., Qin, Y., Chau, M.: Crime data mining: a general framework and some examples. computer **37**(4), 50–56 (2004)
3. Galeotti, M.: Global crime today: the changing face of organised crime. Routledge (2014)
4. Hartwig, M., Anders Granhag, P., Strömwall, L.A.: Guilty and innocent suspects strategies during police interrogations. Psychology, Crime & Law **13**(2) (2007)
5. Hartwig, M., Granhag, P.A., Strömwall, L.A., Vrij, A.: Detecting deception via strategic disclosure of evidence. Law and human behavior **29**(4), 469–484 (2005)
6. Koschade, S.: A social network analysis of jemaah islamiyah: The applications to counterterrorism and intelligence. Studies in Conflict & Terrorism **29**(6) (2006)
7. Lu, Y., Luo, X., Polgar, M., Cao, Y.: Social network analysis of a criminal hacker community. Journal of Computer Information Systems **51**(2), 31–41 (2010)
8. Mason, W., Suri, S.: Conducting behavioral research on amazons mechanical turk. Behavior research methods **44**(1), 1–23 (2012)
9. Paolacci, G., Chandler, J., Ipeirotis, P.G.: Running experiments on amazon mechanical turk. Judgment and Decision making **5**(5), 411–419 (2010)
10. Schwartz, D.M., Rouselle, T.D.: Using social network analysis to target criminal networks. Trends in Organized Crime **12**(2), 188–207 (2009)
11. Sparrow, M.K.: The application of network analysis to criminal intelligence: An assessment of the prospects. Social networks **13**(3), 251–274 (1991)
12. Strömwall, L.A., Willén, R.M.: Inside criminal minds: Offenders' strategies when lying. Journal of Investigative Psychology and Offender Profiling **8**(3) (2011)
13. Suri, S., Watts, D.J.: Cooperation and contagion in web-based, networked public goods experiments. PloS one **6**(3), e16836 (2011)
14. Wijegunawardana, P., Ojha, V., Gera, R., Soundarajan, S.: Sampling dark networks to locate people of interest. Social Network Analysis and Mining **8**(1) (2018)
15. Wong, W.H., Brown, P.A.: E-bandits in global activism: Wikileaks, anonymous, and the politics of no one. Perspectives on Politics **11**(4), 1015–1033 (2013)