Hiding in Social Networks

Summary

Marcin Waniek

Introduction

Recent years have lead to an increased interest in the analysis of social networks [27]. Applications of this wide body of research vary from organizing massive viral-marketing campaigns [7], through the analysis of outbreaks of infectious diseases [11], all the way up to fighting global criminal and terrorist organizations [20, 29].

Clearly, many situations, where social network analysis tools are being applied can be described as adversarial settings. In other words, members of the networks that are being analysed may be interested in falsifying the results of such analysis, e.g., by modifying the structure of their social connections. At the same time, however, nearly all widely used social network analysis tools treat the objects of their analysis as completely oblivious entities.

At the moment we lack sufficient understanding of evasion techniques that can falsify the results of social network analysis tools. What is more, currently existing tools do not even have the ability to internalize such evasion techniques. The reason of this is most social network analysis tools are built around the assumption that members of the network do not act strategically to evade these tools. Even the more advanced tools dedicated to analysing covert networks [26] typically assume that the investigated network is not subject to strategic manipulation. Given this, we want to direct attention towards the strategic evasion of social network analysis tools, as discussed by Michalak et al. [21].

In the dissertation we assume the role of a strategic member (or a group of members) of a social network and investigate whether and how they can evade various social network analysis tools. As tools that we consider take into account only the topology of the network, we assume that possible strategies of members that seek to hide themselves consist of adding and removing edges from the network. We investigate the problem both from the computational complexity point of view, i.e., we ask how hard it is to find an optimal way of hiding, and from a more practical perspective, i.e., we study the settings where, on one hand, either the knowledge or the set of possible strategies of hiding nodes become limited by the external factors and, on the other hand, imperfect, heuristic, easy to implement solutions suffice.

Our work can be considered as a first step in the strategic analysis of hiding in social networks. To the best of our knowledge, we are the first to propose this kind of adversary setting in the social network analysis, where the strategic behaviour of network actors is explicitly considered.
Disguising Centrality of a Node

In Chapter 3 of the dissertation, we analyse the possible ways of avoiding centrality measure analysis. Centrality measure is a function that, using the structure of the network, evaluates the importance of every node. We focus on three fundamental centrality measures, namely the degree centrality, the closeness centrality, and the betweenness centrality, and study how an individual can avoid being identified as a key player by those measures without compromising her influence. Since, from a graph-theoretic perspective, this is fundamentally an optimization problem, we analyse its computational complexity to study the theoretical limits of one’s capability to disguise importance in a social network. Our findings are summarized in Table 1.

<table>
<thead>
<tr>
<th>Problem</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disguising Centrality (Degree)</td>
<td>P</td>
</tr>
<tr>
<td>Disguising Centrality (Closeness)</td>
<td>NP-complete</td>
</tr>
<tr>
<td>Disguising Centrality (Betweenness)</td>
<td>NP-complete</td>
</tr>
<tr>
<td>Influence Recovery (IC)</td>
<td>NP-hard</td>
</tr>
<tr>
<td>Influence Recovery (LT)</td>
<td>NP-hard</td>
</tr>
</tbody>
</table>

Table 1: Summary of our computational hardness results, where LT denotes Linear Threshold influence model and IC denotes Independent Cascade influence model.

Although we show that an optimal solution is hard to compute, we demonstrate the effectiveness of a surprisingly simple heuristic, called ROAM—Remove One, Add Many—whereby the rewiring of social connections is restricted to the individual’s immediate network neighbourhood. Specifically, it involves two actions that are already available on popular social-media platforms, namely “unfriending” a certain friend, and introducing her to a number of our other friends. An illustration of the execution of ROAM is presented in Figure 1. Our results demonstrate, surprisingly, that disguising individuals is in practice possible using a simple, readily implementable heuristic. We quantify the cost-benefit profile of using these heuristics in various empirical networks.

Our main result is that, despite the hardness of finding optimal solutions, disguising oneself can be surprisingly easy in practice, and can be achieved with the use of simple heuristics that can be easily implemented even by lay people.

On one hand, our findings contribute towards charting the limits of protecting privacy in social networks. On the other hand, they expose potentially serious limitations of using generic social network analysis tools in security applications. The fact that such tools can be easily misled underlines the need for developing specialized tools that account for the nature of links and nodes in the network, and not just its topology.

Hiding Leaders

Mapping terrorist networks and analysing their structure is a vitally important part of counter-terrorism efforts. Not only does it help to understand their operational methods
Figure 1: Executing the ROAM heuristic twice on the 9/11 terrorist network to hide Mohamed Atta—one of the ringleaders of the attack [14]. The dotted red edge is the one to be removed by the algorithm, and the dashed green edges are the ones to be added.

and the way of thinking, but also it plays a key role in designing and implementing destabilization strategies [5, 9, 23]. The literature on this research problem agrees that criminals in general, and terrorists in particular, face a trade-off between secrecy and efficiency [24].

In Chapter 4, we propose a theoretical model to study the secrecy-efficiency trade-off that differs from previous ones in a number of ways. Firstly, inspired by studies of real-life covert networks [6, 18], we take a leader-centric approach, i.e., we focus on the role played in a terrorist network by its leadership. In more detail, we investigate how the topology of the network can be deliberately designed to keep the leaders identities hidden. In this context, while the previous literature on identifying leaders of terrorist networks typically assumed that such leaders are not aware of the techniques and methods used by law-enforcement agencies, we assume that this is not the case, i.e., in our model the terrorist leaders strategically shape their network to protect themselves from detection by the centrality measures. In fact, recent media reports and academic studies of criminal and terrorist organizations suggest that members of such organizations are becoming increasingly tech-savvy [25, 12]. Hence, their obliviousness with respect to the available social network analysis techniques should not be taken for granted.
In the first part Chapter 4 we focus on the computational aspects of modifying an existing network so as to shield a group of leaders from centrality analysis. In more detail, we analyse the hardness of rewiring the network so that the ranking of all leaders (based on one of the three main centrality measures) drops below a certain threshold. In comparison with the previous chapter, we now consider a group of nodes that wish to be hidden (while in Chapter 3 we consider only a single node), and we are interested in relative position in centrality ranking (while in Chapter 3 we analysed the complexity of lowering absolute centrality value), and these differences in the setting lead to computational hardness. At first glance, this problem may appear to be easy at least for the degree centrality, which is mathematically uninvolved. Surprisingly, however, we find that this is not the case. The problem of hiding leaders for each of the three centrality measures turns out to be NP-complete. Our results are in line with the literature on modifying a network to increase centrality [8].

Given the hardness of modifying an existing network, we turn out attention to a different question, which is how a covert network can be built from scratch so that the leaders are hidden and, at the same time, have a reasonable influence over the network members. Here the main idea is for the leaders to surround themselves with an “inner circle” of trustees, called “captains”, whose role is to conceal the leaders, and to pass on their communicates to the rest of the network. We identify one such network structure, and prove that every captain is guaranteed to be ranked higher than any of the leaders (according to the three standard centrality measures). Figure 2 illustrates a sample captain network with 3 leaders. In fact, “inner circles” have been identified in various real-life terrorist networks, e.g., in Al-Qaeda [2] and IRA [28]. While we do not have access to data that confirms that those real-life “inner circles” have similar structure to the ones obtained in the dissertation, we hope that our results shed more light on why such circles may exist in covert networks. In this context, charting the topology of covert networks became one of the key research directions.

Hiding Communities

While there exists no precise definition of a community, typically it is understood as a group of nodes who are more densely connected among themselves than with the rest of the network. Various dedicated community detection algorithms have been proposed in the literature. Such algorithms return a community structure, i.e., a partition of the set of nodes into typically-disjoint communities. In this context, the modularity index is a well-known measure of the quality of a particular community structure. As such, any community detection algorithm searches through the space of community structures to select the one that has a particularly high modularity.

In Chapter 5, we investigate whether and how a community could hide itself within a network so that it becomes difficult to identify by various community-detection algorithms. Our motivation is twofold. Firstly, a community may be simply interested in privacy—a value that seems to be increasingly violated by the process of datafication. Consider, for instance, police units that are involved in undercover operations against street gangs. The private life
Figure 2: An illustration of the captain network with 3 leaders. Edges including leaders are depicted as solid black lines; edges between captains are depicted as gray lines; edges between captains and other nodes are depicted as dashed lines.

of this community should not be easily traceable on Facebook or Twitter. What this lack of privacy might lead to has been recently shown by various Ukrainian activists groups who have identified Russian military units in Donbass by analysing social networking sites [1]. As yet another example, staying “below the radar” is imperative for communities of opposition bloggers in authoritarian regimes, because such regimes actively monitor internet content [13].

Our second motivation is related to security. In particular, methods of social network analysis are becoming increasingly used in the fight against criminal organisations [30]. Hence, it is important to understand how various community detection algorithms could be “fooled” by such covert organizations.

We study how a community can conceal itself to increase the likelihood of being overlooked by community-detection algorithms. To this end, we propose a measure of concealment designed to quantify the degree to which a group of individuals is hidden. The measure incorporates two notions of hiding, i.e., distributing members of the group across many communities and hiding in the crowd. Using this measure, we demonstrate the effectiveness of yet another simple heuristic, called DICE—Disconnect Internally, Connect Externally—whereby members of the community either “unfriend” certain other members, or “befriend” some non-members, in order to blend in with the surrounding web of social connections.

**Evading Link Prediction**

One of the key research challenges in social network analysis is the link prediction problem [16, 17]. Intuitively, based on the current structure of the network, this problem involves
predicting the connections that are most likely to be created in the future. An alternative interpretation of this problem is to identify the connections that are hidden from the observer, either due to scarcity of data, or due to the deliberate concealment of information [4].

In Chapter 6, we study how one can make a chosen set of network connections harder to identify. This research question matters because, on one hand, it may assist the general public in protecting their privacy from intrusion by private and public entities; on the other hand, it may mitigate (at least to some extent) the threats posed by cyber criminals. It may also assist law-enforcement agencies in understanding how criminals and terrorists could evade social network analysis tools, especially given their increasing reliance on social-media survival strategies [25, 12].

We show that even for the simplest class of these algorithms, \textit{i.e.}, local similarity indices, the problem of hiding edges is NP-complete. To address this hardness results in a practical way, we propose two polynomial time heuristic solutions. One of them, called OTC—\textit{Open Triad Creation}—focuses on creating as many open triads that do not include edges we intend to hide as possible. The other, called CTR—\textit{Closed Triad Removal}—focuses on removing from the network triads that include edges that we wish to hide. Both heuristics allow to conceal a chosen set of connections by hindering their detection by link prediction algorithms.

\section*{Conclusions}

The following main conclusions can be drawn based on our analysis. Firstly, when considered from the point of view of computational complexity theory, most hiding problems turn out to be NP-hard. Consequently, finding an optimal solution to a given instance is a computationally intensive task, and cannot be effectively solved for most networks. In many cases even for simplest social network analysis tools, such as degree centrality and common neighbours similarity index, considered problems prove to be intractable. It suggests that hardness of the hiding questions is the result of a structure of the problem itself, and not only of the complexity of the social network analysis tools that we try to evade. Since in all considered settings we aimed to use the most basic tools (that can be used even to analyse massive social networks), we expect the problems of hiding from more complex social network analysis techniques to be even less tractable.

Secondly, even though the conclusion drawn in the previous paragraph may suggest that the network analysers have nothing to be worried about, this is certainly not the case. For every considered problem we managed to find a simple polynomial-time heuristic solution that turned out to be effective in practice, for both artificial and real-life datasets. Each of these algorithms takes into account several factors that are crucial while developing solutions that are implementable in practice. Our heuristic solutions do not need complete knowledge about entire structure of the network, as such information is rarely available in practice. Typically, we only know the structure of our direct network vicinity, \textit{i.e.}, connections of our neighbours and sometimes neighbours of our neighbours. Our heuristic algorithms do not require extensive computations, \textit{i.e.}, they are all polynomial time (for low-ordered polynomials). As real-life network can be massive, it is important to develop solutions that do not require
extensive computational power. Most of our heuristic solutions are also simple enough that 
they can be used even by lay people, without specialized knowledge about algorithm design. 
Consequently, they can be implemented by any social media users. All things considered, 
even though finding the optimal solutions to the hiding problems is a very hard task from a 
computational point of view, coming up with a solution that provides an acceptable level of 
concealment is often relatively straightforward.

Third conclusion that can be drawn from the dissertation is the need to extend the ap-
proaches used to analyse social networks, especially when considering the dark networks. 
Nearly all social network analysis tools treat the nodes of the network as oblivious entities. 
While this might be true in many cases, when considering criminal and terrorist organiza-
tions, their members have clear incentive to falsify the results of the analysis. As we have 
shown in the dissertation, this goal can be achieved even with relatively simple means, e.g., 
by modifying the structure of connections between members of the organization or by re-
organizing its communication channels. This indicates, that when performing the analysis 
on this particular type of networks, nodes should be considered as strategic players. Since 
today’s criminal and terrorist organizations are aware of being under constant surveil-
ance, they can modify their behaviour in order to minimize the damage. This possibility should 
be brought to attention of police forces and intelligence agencies utilizing social network 
analysis tools.

Fourth conclusion is the need to develop new social network analysis tools, that are aware 
of the potential attempts to avoid and falsify the analysis. The dissertation describes flaws 
of many of the currently used techniques and risks that are the result of commonly accepted 
assumptions. Next generations of social network analysis tools should adapt to the fact that 
their object of analysis is not static and unaware of their existence. One of possible ways 
to achieve this, is developing techniques taking into account more than just the structure 
of connections between network members, but also specific information about each of them. 
While current software is usually just visualising this type of data, its usage should be built 
into analytic algorithms. This may lead to more effective analysis of dark networks and, as 
a result, more effective ways of fighting criminal and terrorist organizations.

References


[2] V. Barber. The evolution of al qaeda’s global network and al qaeda core’s position 
within it: A network analysis. Perspectives on Terrorism, 9(6), 2015.


