Hiding in Multilayer Networks

Marcin Waniek,[1,2] **Tomasz P. Michalak,**[2] **Talal Rahwan**[1]
[1]Computer Science, New York University Abu Dhabi, [2]Institute of Informatics, University of Warsaw
mjwaniek@gmail.com, tpm@mimuw.edu.pl, tr72@nyu.edu

Abstract

Multilayer networks allow for modeling complex relationships, where individuals are embedded in multiple social networks at the same time. Given the ubiquity of such relationships, these networks have been increasingly gaining attention in the literature. This paper presents the first analysis of the robustness of centrality measures against strategic manipulation in multilayer networks. More specifically, we consider an "evader" who strategically chooses which connections to form in a multilayer network in order to obtain a low centrality-based ranking—thereby reducing the chance of being highlighted as a key figure in the network—while ensuring that she remains connected to a certain group of people. We prove that determining an optimal way to "hide" is NP-complete and hard to approximate for most centrality measures considered in our study. Moreover, we empirically evaluate a number of heuristics that the evader can use. Our results suggest that the centrality measures that are functions of the entire network topology are more robust to such a strategic evader than their counterparts which consider each layer separately.

Introduction

Owing to several incidents in the past few years, most notably those concerning the American presidential elections of 2016, the general public has become increasingly concerned with the privacy and security of their online activities (Persily 2017). Experts, however, had been warning about such potential risks long ago. For instance, Mislove et al. (2010) famously showed that, by coupling the social network of a given Facebook user with publicly-known attributes of some other users, it is possible to infer otherwise-private information about that user. Worryingly, this is true not only for typically innocuous data, but also for potentially-sensitive confidential information such as political preferences (as demonstrated in the case of Cambridge Analytica), or even sexual orientation (Kitchin 2016).

Various proposals on how to deal with such privacy challenges have already been put forward. Among those proposals is the General Data Protection Regulation, implemented in May 2018, which is perhaps the most well-known attempt to use state-enforced, legal instruments (EU 2016). On the other hand, there have been a plethora of algorithmic solutions for privacy protection (Lane et al. 2014; Kearns et al. 2016). Perhaps the most well-known such solutions come from the network anonymization and de-anonymization literature (Zhou, Pei, and Luk 2008; Narayanan and Shmatikov 2009; Kayes and Iamnitchi 2015), which studies the problem faced by a *data trustee* who publishes anonymized network data to be analyzed for various purposes. In this literature, the responsibility of protecting the privacy of the network members lies solely on the shoulders of the data trustee, while the network members are implicitly assumed to be passive in this regard. In contrast, a recent body of work studies ways in which the network members can themselves protect their own privacy by acting strategically to evade various tools from the social network analysis toolkit (Michalak, Rahwan, and Wooldridge 2017). In this context, three fundamental classes of tools have been considered: (1) centrality measures, (2) community detection algorithms; and (3) link prediction algorithms. More specifically, Waniek et al. (2018; 2017) studied how key individuals in a social network could rewire the network to avoid being highlighted by centrality measures while maintaining their own influence within the network. The authors also studied how a group of individuals could avoid being identified by community detection algorithms. Furthermore, Yu et al. (2018), Waniek et al. (2019), and Zhou et al. (2019) studied how to hide one's sensitive relationships from link prediction algorithms.

The aforementioned literature on the strategic behaviour of network members demonstrates that it is indeed possible to develop reasonably effective heuristics to escape detection by fundamental network analysis tools. However, the main limitation of this literature is that it focuses only on standard, single-layered networks. In contract, people often interact with each other via a complicated pattern of relationships, thereby creating multiple subsystems, or "layers", of connectivity. This is even more so nowadays when many of us belong to multiple social media platforms simultaneously. Furthermore, multilayer networks are increasingly being recognized not only in the context of human interactions, but also in many natural and engineered systems (De Domenico et al. 2013). For instance, to travel from one point to another in many urban transportation networks, one can choose between a road subnetwork (car or taxis), bus or tram subnetwork, subway subnetwork, local train

subnetwork, bike subnetwork, footpath subnetwork, or any combination thereof. Each such subnetwork has its own distinct characteristics, which become difficult, or even impossible, to account for if modelled as a single layer due to the interdependencies between the different layers. The theoretical and empirical analysis of multilayer networks has recently attracted significant attention (see the work by Kivelä et al. (2014) for a comprehensive review). This new body of research is primarily driven by the fact that, due to the much more complex nature of multilayer networks, many results for singlelayer networks become obsolete.

Motivated by these observations, we present in this paper the first analysis of how to protect ones' privacy against centrality measures in multilayer networks. Specifically, we consider an evader who wishes to connect to a certain group of individuals, without being highlighted by centrality measures as a key member in the multilayer network. To this end, the evader has to strategically choose at which layer(s) to connect to those individuals. We prove that the corresponding optimization problem is NP-complete and hard to approximate for most centrality measures considered in our study. Furthermore, we empirically evaluate a number of heuristic algorithms that the evader can use. The results of this evaluation suggest that the centrality measures that are functions of the entire network topology are more robust to such a strategic evader than their counterparts which consider each layer separately.

Preliminaries

Basic Network Notation and Definitions: Let $G = (V, E)$ denote a *simple (single-layer) network*, where V is the set of n nodes and $E \subseteq V \times V$ the set of edges. We denote an edge between nodes v and w by (v, w).

In this paper we consider *multilayer networks*, i.e., networks where edges can represent different types of relations. We will denote a multilayer network by $M = (V_L, E_L, V, L) \in \mathbb{M}$, where V is the set of nodes, L is the set of layers (i.e., types of relations), $V_L \subseteq V \times L$ is the set of occurrences of nodes in layers (e.g., having $(v, \alpha) \in V_L$ means that node v appears in layer α), and $E_L \subseteq V_L \times V_L$ is the set of edges. We will denote an occurrence of node v in layer α by v^α. Note that $V = \{v : \exists_{\alpha \in L} v^\alpha \in V_L\}$. Let V^α be the set of nodes occurring in layer α, i.e., $V^\alpha = \{v \in V : v^\alpha \in V_L\}$, and let G^α denote the simple network consisting of all the nodes and edges in layer α, i.e., $G^\alpha = (V^\alpha, \{(v, w) : (v^\alpha, w^\alpha) \in E_L\})$.

We focus on *undirected* networks, i.e., we do not discern between edges (v^α, w^β) and (w^β, v^α). Moreover, we do not consider self-loops, i.e., $\forall_{v^\alpha \in V_L}(v^\alpha, v^\alpha) \notin E_L$. Multilayer network allow for *inter-layer edges*, which are edges between two layers; they may connect two different nodes, or may connect two occurrences of the same node. We restrict our attention to networks with *diagonal couplings*, i.e., networks where every inter-layer edge connects two occurrences of the same node, i.e., $\forall_{(v^\alpha, w^\beta) \in E_L} \alpha \neq \beta \rightarrow v = w$.

Notice that, in some literature, multilayer networks with diagonal couplings are called *multiplex* networks. However, it is also typically assumed that the multiplex networks are node-aligned (i.e., every node occurs in every layer), which is not the case in our setting. Hence, we will use the more general term "*multilayer networks*". For a comprehensive discussion of the nomenclature, see Kivelä et al. (2014).

A path in a simple network is an ordered sequence of nodes in which every two consecutive nodes are connected by an edge. A path in a multilayer network is an ordered sequence of node occurrences in which every two consecutive occurrences are connected by an edge. The length of a path is the number of edges in that path. The set of all shortest paths between a pair of nodes, $v, w \in V$ will be denoted by $\pi_G(v, w)$. The distance between a pair of nodes $v, w \in V$ is the length of a shortest path between them, and is denoted by $\lambda_G(v, w)$. We assume that if there does not exist a path between v and w then $\lambda_G(v, w) = \infty$. In a multilayer network we consider distance between v and w to be the shortest distance between an occurrence of v in any layer α and an occurrence of w in any layer β (possibly $\alpha \neq \beta$).

For any node, $v \in V$, in a simple network, G, we denote by $N_G(v) = \{w \in V : (v, w) \in E\}$ the set of neighbors of v in G. Similarly, given a multilayer network M, we write $N_M(v) = \{w \in V : (v^\alpha, w^\beta) \in E_L\}$. Finally, we denote by $N_M^\alpha(v)$ the set of neighbors of v in layer α, i.e., $N_M^\alpha(v) = \{w \in V : (v^\alpha, w^\alpha) \in E_L\}$. We will often omit the network itself from the notation whenever it is clear from the context, e.g., by writing $\lambda(v, w)$ instead of $\lambda_G(v, w)$.

Centrality Measures: A centrality measure (Bavelas 1948) is a function that expresses the importance of a given node in a given network. Arguably, the best-known centrality measures are *degree*, *closeness* and *betweenness*.

Degree centrality (Shaw 1954) assumes that the importance of a node is proportional to the number of its neighbors, i.e., the degree centrality of node v in network G is:

$$c_{degr}(G, v) = |N_G(v)|.$$

Closeness centrality (Beauchamp 1965) quantifies the importance of a node in terms of shortest distances from this node to all other nodes in the network. Formally, the closeness centrality of node v in network G can be expressed as:

$$c_{clos}(G, v) = \sum_{w \in V \setminus \{v\}} \frac{1}{\lambda_G(v, w)}.$$

Betweenness centrality (Anthonisse 1971; Freeman 1977) states that, if we consider all the shortest paths in the network, then the more such paths traverse through a given node (it is often stated that the node *controls* such paths), the more important the role of that node in the network. More formally, the betweenness centrality of node $v \in V$ in network G is:

$$c_{betw}(G, v) = \sum_{w, u \in V \setminus \{v\}} \frac{|\{p \in \pi_G(w, u) : v \in p\}|}{|\pi_G(w, u)|}.$$

The definitions of degree and closeness centrality can be generalized to multilayer networks using the definitions of neighbors and distance for multilayer networks (see above). As for the betweenness centrality of node v in a multilayer network M, it grows with the number of occurrences of v

on the shortest paths between pairs of other nodes:

$$c_{betw}(M,v) = \sum_{w,u \in V \setminus \{v\}} \frac{|\{(v^\alpha, p) : v^\alpha \in p, p \in \pi_M(w,u)\}|}{|\pi_M(w,u)|}$$

To avoid any potential confusion, the measures that are designed for simple networks will be referred to as "*local* centrality measures", since they can be applied to only a single layer. Conversely, the measures that are designed for multilayer networks will be referred to as a "*global* centrality measures", since they take all layers into consideration.

Theoretical Analysis

In this section we formally define our computational problems and then move on to analyse them.

Definitions of Computational Problems

We define the decision problems before defining the corresponding optimization problems. Here, the "group of contacts" is the set of individuals to whom the evader wishes to connect while remaining hidden from centrality measures.

Decision Problems: We will define two different decision versions of this problem, starting with the global version.

Definition 1 (Multilayer Global Hiding)**.** *This problem is defined by a tuple, $(M, \widehat{v}, F, c, d)$, where M is a multilayer network $M = (V_L, E_L, V, L)$, $\widehat{v} \in V$ is the evader, $F \subset V$ is the group of contacts, c is a centrality measure, and $d \in \mathbb{N}$ is a safety margin. The goal is to identify a set of edges to be added to the network, $A^* \subseteq \{(\widehat{v}^\alpha, v^\alpha) : v \in F \wedge \widehat{v}^\alpha \in V_L \wedge v^\alpha \in V_L\}$, such that in the resulting network $\widehat{M} = (V_L, E_L \cup A^*, V, L)$ the evader is connected with every contact in at least one layer and there are at least d nodes with a centrality score greater than that of the evader, i.e.:*

$$\forall_{v \in F} \exists_{\alpha \in L} (\widehat{v}^\alpha, v^\alpha) \in A^*,$$

$$\exists_{W \subset V} \left(|W| \geq d \wedge \forall_{v \in W} c(\widehat{M}, v) > c(\widehat{M}, \widehat{v}) \right).$$

We say that "$\widehat{v}$ is hidden" when there are at least d nodes whose centrality is greater than that of $\widehat{v}$.

Definition 2 (Multilayer Local Hiding)**.** *This problem is defined by a tuple, $(M, \widehat{v}, F, c, (d^\alpha)_{\alpha \in L})$, where $M = (V_L, E_L, V, L)$ is a multilayer network, $\widehat{v} \in V$ is the evader, $F \subset V$ is the group of contacts, c is a centrality measure, and $d^\alpha \in \mathbb{N}$ is a safety margin for layer $\alpha \in L$. The goal is to identify a set of edges to add, $A^* \subseteq \{(\widehat{v}^\alpha, v^\alpha) : v \in F \wedge \widehat{v}^\alpha \in V_L \wedge v^\alpha \in V_L\}$, such that in the resulting network $\widehat{M} = (V_L, E_L \cup A^*, V, L)$ the evader is connected with every contact in at least one layer and for each layer α the network G^α contains at least d^α nodes with a centrality score greater than that of the evader, i.e.:*

$$\forall_{v \in F} \exists_{\alpha \in L} (\widehat{v}^\alpha, v^\alpha) \in A^*,$$

$$\forall_{\alpha \in L} \exists_{W \subset V^\alpha} \left(|W| \geq d^\alpha \wedge \forall_{v \in W} c(\widehat{M}^\alpha, v) > c(\widehat{M}^\alpha, \widehat{v}) \right).$$

We say that "$\widehat{v}$ is hidden in α" if there are at least d^α nodes whose centrality in layer α is greater than that of $\widehat{v}$ in α.

In the global version of the problem we assume that the seeker is able to observe and analyze the entire multilayer network using centrality measures, hence the evader's goal is to minimize her centrality ranking in the network as a whole. On the other hand, the local version of the problem models situations where the seeker analyzes only one of the layers, e.g., if the seeker gains access to the email communication network, but not to the phone-call network. In such situations, the evader's goal is to attain an adequate level of safety in each layer separately.

The approach to hiding represented by the two problems differs from the one developed for simple networks by Waniek et al. (2017; 2018). Their hiding algorithms focus on choosing which edge(s) to add or remove from the single layer, often causing the evader to lose the direct connection with some of the neighbors. The algorithms presented in our paper focus on choosing the layer in which to maintain the connection, and allow the evader to keep direct links with all contacts. Notice that this approach cannot be applied to simple networks, as there is only one way to have a direct link between the evader and every contact in a single layer.

Optimization Problems: We now define the corresponding optimization problems. They take into consideration a situation when it is impossible to connect the evader with all the contacts.

Definition 3 (Maximum Multilayer Global Hiding)**.** *This problem is defined by a tuple, $(M, \widehat{v}, F, c, d)$, where $M = (V_L, E_L, V, L)$ is a multilayer network, $\widehat{v} \in V$ is the evader, $F \subset V$ is the group of contacts, c is a centrality measure, and $d \in \mathbb{N}$ is a safety margin. The goal is then to identify a set of edges to be added to the network, $A^* \subseteq \{(\widehat{v}^\alpha, v^\alpha) : v \in F \wedge \widehat{v}^\alpha \in V_L \wedge v^\alpha \in V_L\}$, such that in the resulting network $\widehat{M} = (V_L, E_L \cup A^*, V, L)$ the evader is connected with as many contacts as possible, while there are at least d nodes with a centrality score greater than that of the evader.*

Definition 4 (Maximum Multilayer Local Hiding)**.** *This problem is defined by a tuple, $(M, \widehat{v}, F, c, (d^\alpha)_{\alpha \in L})$, where $M = (V_L, E_L, V, L)$ is a multilayer network, $\widehat{v} \in V$ is the evader, $F \subset V$ is the group of contacts, c is a centrality measure, and $d^\alpha \in \mathbb{N}$ is a safety margin for layer $\alpha \in L$. The goal is then to identify a set of edges to be added to the network, $A^* \subseteq \{(\widehat{v}^\alpha, v^\alpha) : v \in F \wedge \widehat{v}^\alpha \in V_L \wedge v^\alpha \in V_L\}$, such that in the resulting network $\widehat{M} = (V_L, E_L \cup A^*, V, L)$ the evader is connected with as many contacts as possible, while for each layer α the network G^α contains at least d^α nodes with a centrality score greater than that of the evader.*

Intuitively, the goal is to connect the evader with as many contacts as possible, while keeping the evader hidden.

Complexity Analysis

The complexity results for both the global and local versions of the problem are listed below (see Table 1 for a summary). Due to space constraints, we present here only the proof of Theorem 5; all the remaining proofs can be found in Waniek et al. (2019).

Table 1: Summary of our computational complexity results.

Centrality	Multilayer Global Hiding	Multilayer Local Hiding
Degree	P	NP-complete
Closeness	NP-complete	NP-complete
Betweenness	NP-complete	NP-complete

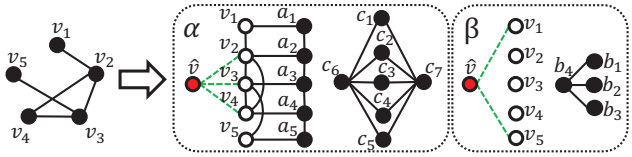

Figure 1: An illustration of the network used in the proof of Theorem 5. The red node represents the evader, while the white nodes represent the contacts. Dashed (green) edges represent the solution to this problem instance.

Observation 1. *The problem of Multilayer Global Hiding is in P given the degree centrality measure. In fact, for a given problem instance either any A^* that connects $\widehat{v}$ with all contacts is a solution, or there are no solutions at all.*

Theorem 1. *The problem of Multilayer Global Hiding is NP-complete given the closeness centrality measure.*

Theorem 2. *The problem of Multilayer Global Hiding is NP-complete given the betweenness centrality measure.*

Theorem 3. *The problem of Multilayer Local Hiding is NP-complete given the degree centrality measure.*

Theorem 4. *The problem of Multilayer Local Hiding problem is NP-complete given the closeness centrality measure.*

Theorem 5. *The problem of Multilayer Local Hiding is NP-complete given the betweenness centrality measure.*

Proof of Theorem 5: The problem is trivially in NP, since after the addition of a given A^* the betweenness centrality rankings for all layers can be computed in polynomial time.

Next, we prove that the problem is NP-hard. To this end, we show a reduction from the NP-complete problem of *Finding k-Clique*. The decision version of this problem is defined by a simple network, $G = (V, E)$, and a constant, $k \in \mathbb{N}$. The goal is then to determine whether there exist k nodes in G that form a clique.

Let us assume that $k < n - 1$ (if this assumption does not hold then the solution can be computed in polynomial time). Furthermore, let us assume that G is connected (if this does not hold, the problem can be considered separately for each connected component). Given an instance of the problem of *Finding k-Clique* where $k < n - 1$, and given a simple network $G = (V, E)$, let us construct a multilayer network, $M = (V_L, E_L, V', L)$, as follows (Figure 1 depicts an instance of this network):

- **The set of nodes V':** This consists of the following sets of nodes: $V = \{v_1, \ldots, v_n\}$, $A = \{a_1, \ldots, a_n\}$, $B = \{b_1, \ldots, b_{n-k+2}\}$, and $C = \{c_1, \ldots, c_{n+2}\}$.
- **The set of layers L:** We create two layers, α and β.
- **The set of occurrences of nodes in layers V_L:** Layer α contains all nodes in $\{\widehat{v}\} \cup V \cup A \cup C$, while layer β contains all nodes in $\{\widehat{v}\} \cup V \cup B$.
- **The set of edges E_L: In layer** α we create an edge between two nodes $v_i, v_j \in V$ if and only if this edge was present in G. We also create an edge (v_i, a_i) for every v_i, and an edge between every pair a_i, a_{i+1}. Finally, for every node $c_i \in C : i \leq n$, we create edges (c_i, c_{n+1}) and (c_i, c_{n+2}). **In layer** β we create an edge (b_i, b_{n-k+2}) for every node $b_i \in B : i < n - k + 2$.

Now, consider the following instance of the problem of Multilayer Local Hiding, $(M, \widehat{v}, F, c, (d^\alpha)_{\alpha \in L})$, where: M is the multilayer network we just constructed; $\widehat{v}$ is the evader; $F = V$ is the set of contacts; c is the betweenness centrality measure; and $d^\alpha = 3n + 2$ and $d^\beta = 1$ are the safety margins. Given this, let us consider what are the sets of edges that can be added between the evader $\widehat{v}$ and the contacts F in each layer, so that the evader is hidden.

Since $d^\alpha = 3n + 2$, then apart from the evader $\widehat{v}$, the betweenness centrality of every node in layer α must be greater than that of $\widehat{v}$; otherwise the evader $\widehat{v}$ would not be hidden in α. Also note that the betweenness centrality of every node $c_i \in C : i \leq n$ equals $\frac{1}{n}$, and all nodes other than $\widehat{v}$ have non-zero betweenness centrality.

Now if $\widehat{v}$ gets connected to any two nodes $v_i, v_j \in V$ that are not connected to one another, then $\widehat{v}$ controls one shortest path of length 2 between v_i and v_j. Note that there can be at most $n-2$ other shortest paths of length 2 between v_i and v_j (each such path goes through some node $v_k \in V \setminus \{v_i, v_j\}$ if and only if v_k is connected to both v_i and v_j). Thus, the betweenness centrality of $\widehat{v}$ is at least $\frac{1}{n-1}$. Consequently, all nodes that $\widehat{v}$ is connected to in layer α must form a clique in order for $\widehat{v}$ to be hidden in α.

Consider a situation in which the evader $\widehat{v}$ is connected to x nodes from V in layer β (notice that $x \leq n$). Its betweenness centrality is then $\frac{x(x-1)}{2}$, as it controls all shortest paths between pairs of its neighbors, but not any other shortest paths. At the same time, the betweenness centrality of the node b_{n-k+2} is $\frac{(n-k+1)(n-k)}{2}$ (as it controls all shortest paths between pairs of other nodes from B), which is greater than the betweenness centrality of $\widehat{v}$ if and only if $x \leq n-k$. All other nodes in the layer have betweenness centrality 0. Thus, $\widehat{v}$ is hidden in β iff it has at most $n - k$ neighbors.

Now we will show that if there exists a solution to the given instance of the problem of Finding k-Clique, then there also exists a solution to the constructed instance of the problem of Multilayer Local Hiding. To this end, let V^* be a group of k nodes forming a clique in G. Let us create A^* by connecting $\widehat{v}$ to nodes from V^* in layer α and to nodes from $F \setminus V^*$ in layer β. As argued above, for such A^*, the evader $\widehat{v}$ is hidden in both layers, hence A^* is a solution to the constructed instance of the Multilayer Local Hiding problem.

To complete the proof we have to show that if there exists a solution A^* to the constructed instance of the problem of Multilayer Local Hiding, then there also exists a solution to the given instance of the problem of Finding k-Clique. Since $\widehat{v}$ can be connected in layer β to at most $n - k$ nodes from V, it has to have at least k neighbors from V in layer α

As shown above, in order for $\widehat{v}$ to be hidden in α, all of its neighbors must form a clique. Hence, the neighbors of $\widehat{v}$ in layer α form a clique in G. This concludes the proof. □

Approximation Analysis

In this section we present the analysis of optimization versions of our problems (see Table 2 for a summary). Again, due to space constraints, we present only the proof of Theorem 9; all remaining proofs are in Waniek et al. (2019).

Theorem 6. *The Maximum Multilayer Global Hiding problem can be solved in polynomial time.*

Theorem 7. *Maximum Multilayer Global Hiding problem given the closeness centrality cannot be approximated within $|F|^{1-\epsilon}$ for any $\epsilon > 0$, unless P=NP.*

Theorem 8. *Both Maximum Multilayer Global Hiding and Maximum Multilayer Local Hiding problems given the betweenness centrality cannot be approximated within $|F|^{1-\epsilon}$ for any $\epsilon > 0$, unless P=NP.*

Theorem 9. *The greedy algorithm is a 2-approximation for the Maximum Multilayer Local Hiding problem given the degree centrality. The bound is tight.*

Proof of Theorem 9: First, let us analyze the structure of a solution to the Maximum Multilayer Local Hiding problem given the degree centrality. Let δ^α be the degree of the d^α-th node in the degree centrality ranking of the nodes in V^α, let δ_0^α be the initial (i.e., before any edges to the contacts are added) degree of the evader in layer α, and let F^α be the set of occurrences of contacts in layer α, i.e., $F^\alpha = \{v^\alpha : v \in F\}$. An algorithm solving the Maximum Multilayer Local Hiding problem can either:

a) connect the evader to at most $k^\alpha = \delta^\alpha - 1 - \delta_0^\alpha$ of freely selected nodes from F^α, as this way the degree of the evader is increased to at most $\delta^\alpha - 1$, and the nodes from the first d^α positions of the degree ranking before the addition continue to have greater degree than the evader when the new edges are added;

b) connect the evader to exactly $\delta^\alpha - \delta_0^\alpha$ nodes from F^α (notice that $\delta^\alpha - \delta_0^\alpha = k^\alpha + 1$). This increases the degree of the evader to δ^α, hence the new connections must include at least $d^\alpha - |\{v^\alpha \in V^\alpha : |N^\alpha(v)| > \delta^\alpha\}|$ nodes with degree exactly δ^α. As a result, there will now exist d^α nodes with degree at least $\delta^\alpha + 1$ and the safety margin will be maintained.

First, notice that the sets of potential connections in both a) and b) can be easily computed in polynomial time, hence the greedy algorithm can use them to optimize the choice of edges added in a single layer.

Notice also that the evader can never add more than $k^\alpha + 1$ edges in layer α, as her degree will then increase to at least $\delta^\alpha + 2$. Since adding a set of connections between the evader and the contacts cannot increase the degree of any contact by more than one, the d^α-th node in the degree centrality ranking of the nodes in V^α will have degree at most $\delta^\alpha + 1$. Hence, the safety margin cannot be maintained.

Finally, notice that if $k^\alpha < 0$, then the degree of the evader is at least δ^α before adding any edges, which puts her within the top d^α positions of the degree centrality ranking. Since increasing the degree of any other nodes can be realized only by adding an edge to the evader (which in turn increases the evader's degree even more), the problem does not have a solution if $k^\alpha < 0$ for any layer α.

The greedy algorithm iterates over the layers and for each layer it connects the evader with maximum possible number of contacts that the evader has not been connected with yet. Notice that it is never beneficial to connect the evader with a given contact in more than one layer, hence any solution doing so has an equivalent solution without the redundant edge(s). In what follows, we will only consider solutions without the redundant edges.

Let us now compare a solution $A^\$$ returned by the greedy algorithm with an optimal solution A^*. We will denote by $A^\$_\alpha$ the set of contacts connected to the evader by the greedy algorithm in layer α, i.e., $A^\$_\alpha = \{v \in V^\alpha : (\widehat{v}^\alpha, v^\alpha) \in A^\$\}$, and by A^*_α the set of contacts connected to the evader by the optimal algorithm in layer α, i.e., $A^*_\alpha = \{v \in V^\alpha : (\widehat{v}^\alpha, v^\alpha) \in A^*\}$. We iterate over the layers of the network in the same order as the greedy algorithm; let this order be $\alpha_1, \ldots, \alpha_{|L|}$. Contacts that the optimal solution connects the evader to in a given layer α_i can be grouped into three pairwise disjoint sets:

- Contacts that could not have been selected in layer α_i by the greedy algorithm, as they were selected by it in one of the previous layers, i.e.:

$$X^{\alpha_i} = \{v \in A^*_{\alpha_i} : v \notin A^\$_{\alpha_i} \wedge \exists_{j<i} v \in A^\$_{\alpha_j}\};$$

- Contacts that are not selected by the greedy algorithm in layer α_i, but they could have been selected, i.e.:

$$Y^{\alpha_i} = \{v \in A^*_{\alpha_i} : v \notin A^\$_{\alpha_i} \wedge \neg\exists_{j<i} v \in A^\$_{\alpha_j}\};$$

- Contacts that are selected by both the greedy algorithm and the optimal solution in layer α_i, i.e.:

$$Z^{\alpha_i} = \{v \in A^*_{\alpha_i} : v \in A^\$_{\alpha_i}\}.$$

We will show that $|A^*_{\alpha_i}| - |A^\$_{\alpha_i}| \leq |X^{\alpha_i}|$, i.e., the difference between the number of edges added in layer α_i by the optimal solution and by the greedy algorithm cannot be greater than $|X^{\alpha_i}|$. We will prove this by contradiction. To this end, assume that in some layer α_i the said difference is greater than $|X^{\alpha_i}|$, i.e., $|A^*_{\alpha_i}| > |A^\$_{\alpha_i}| + |X^{\alpha_i}|$. Since $|A^*_{\alpha_i}| = |X^{\alpha_i}| + |Y^{\alpha_i}| + |Z^{\alpha_i}|$, we get that in this layer: $|A^\$_{\alpha_i}| < |Y^{\alpha_i}| + |Z^{\alpha_i}|$. However, since none of the nodes from $Y^{\alpha_i} \cup Z^{\alpha_i}$ were selected by the greedy algorithm in the previous layers, the greedy algorithm would have chosen to connect the evader with contacts from $Y^{\alpha_i} \cup Z^{\alpha_i}$, as it connects the evader with a greater number of nodes in layer α_i than the solution $A^\$$. Therefore, the difference between the number of edges added in layer α_i by the optimal solution and by the greedy algorithm cannot be greater than $|X^{\alpha_i}|$, i.e., $|A^*_{\alpha_i}| - |A^\$_{\alpha_i}| \leq |X^{\alpha_i}|$. Summing over all layers yields:

$$\sum_{\alpha_i \in L} |A^*_{\alpha_i}| \leq \sum_{\alpha_i \in L} |A^\$_{\alpha_i}| + \sum_{\alpha_i \in L} |X^{\alpha_i}|.$$

Table 2: Summary of our results regarding approximation algorithms.

Centrality	Maximum Multilayer Global Hiding	Maximum Multilayer Local Hiding
Degree	can be solved in polynomial time	greedy algorithm is 2-approximation
Closeness	-	cannot be approximated within $\|F\|^{1-\epsilon}$ for any $\epsilon > 0$
Betweenness	cannot be approximated within $\|F\|^{1-\epsilon}$ for any $\epsilon > 0$	cannot be approximated within $\|F\|^{1-\epsilon}$ for any $\epsilon > 0$

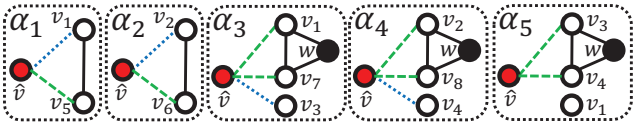

Figure 2: An illustration of the network showing the tightness of the bound given in Theorem 9. The red node represents the evader, while white the nodes represent the contacts. Dashed (green) edges represent the optimal solution to this problem instance, while dotted (blue) edges represent the solution returned by the greedy algorithm.

Since any v is a member of only a single set X^{α_i} (as we assumed that the optimal solution does not contain any redundant edges) and since from the definition of X^{α_i} we have that $\exists_{j<i} v \in A^{\$}_{\alpha_j}$, we get that $\sum_{\alpha_i \in L} |X^{\alpha_i}| \leq |A^{\$}|$. Given that $\sum_{\alpha_i \in L} |A^*_{\alpha_i}| = |A^*|$ and $\sum_{\alpha_i \in L} |X^{\alpha_i}| = |A^{\$}|$ we get:

$$|A^*| \leq 2|A^{\$}|.$$

Since we consider solution without redundant edges, the size of each solution is equal to the number of contact connected with the evader by each solution. Therefore, the greedy algorithm is a 2-approximation.

Figure 2 presents an example of the network, where the bound is tight, i.e., the optimal solution connects the evader with exactly twice as many contacts as the greedy algorithm. The green edges represent the optimal solution, connecting the evader with all eight contacts, while the greedy algorithm (the result of which is represented by the blue edges) connects the evader to only four contacts. □

Heuristics & Empirical Analysis

Given that most computational results are negative, we shift now our attention towards developing heuristic algorithms that provide efficient, albeit not optimal, solutions.

Heuristic Algorithms

Recall that the "group of contacts" refers to the set of individuals whom the evader wishes to connect to. We will refer to each member of this group as a "contact". Notice that a typical member of a social network does not have complete knowledge about the network's structure. Hence, we assume that the evader's knowledge is limited to the connections between the contacts, as well as the degree of each contact. All of our heuristic algorithms take only this information into account. Specifically:

- *Random*—This heuristic connects the evader to every contact in a layer chosen uniformly at random out of all layers in which both the evader and that contact occur.
- *All in one*—This heuristic (Algorithm 1) focuses on creating edges between the evader and her contacts in as few layers as possible. The intuition is that, by focusing all activities of the evader in a small number of layers (if possible, in only one layer), the global centrality measures would assign low importance to the evader. Even though this heuristic might seem overly simplified, we include it as a reasonable baseline—a "rule of thumb" that could be readily implemented by members of the general public.
- *Fringe*—This heuristic (Algorithm 2) focuses on minimizing the number of nodes that are in close vicinity of the evader. The main idea behind this heuristic is to maximize the average distance between the evader and other nodes, in the hope of achieving low ranking according to closeness centrality. Given the limited knowledge of the evader about the network topology, the heuristic cannot analyze any nodes whose distance from the evader is greater than 2. Therefore, the heuristic simply focuses on minimizing the number of neighbors of the contacts.
- *Density*—This heuristic (Algorithm 3) is meant to link the evader to densely connected groups in each layer. Here, the underlying idea is that edges between the contacts act as "shortcuts", preventing the shortest paths in the network from running through the evader, thus reducing her betweenness centrality. Intuitively, the heuristic prefers to connect the evader to a contact v in layers where v is connected to many nodes that are already connected to the evader (the term $|\{w \in F : (\widehat{v}^{\alpha}, w^{\alpha}) \in A\} \cap N^{\alpha}(v)|$ in the numerator), as well as layers where v has many connections with other contacts (the term $|F \cap N^{\alpha}(v)|$ in the numerator) to increase the chance of creating additional "shortcuts". Finally, the heuristic prefers layers with fewer contacts connected to the evader (the term $|\{w \in F : (\widehat{v}^{\alpha}, w^{\alpha}) \in A\}|$) to distribute the evader's connections among layers more uniformly, thereby helping her hide from local centrality measures.

The Simulation Process

In our simulations, we consider *local* degree, closeness and betweenness centrality, as well as *global* closeness and betweenness centrality. The reason behind excluding global degree centrality is that, as stated in Observation 1, for any given group of contacts, the centrality ranking of the evader does not depend on the way in which connections are distributed across the different layers. The simulation process is as follows. For every network, we pick as potential evaders the nodes that are ranked among the top 10 according to at least one of the five considered centrality measures. We then simulate the hiding process for each one of those evaders separately. To this end, we choose the group of contacts to

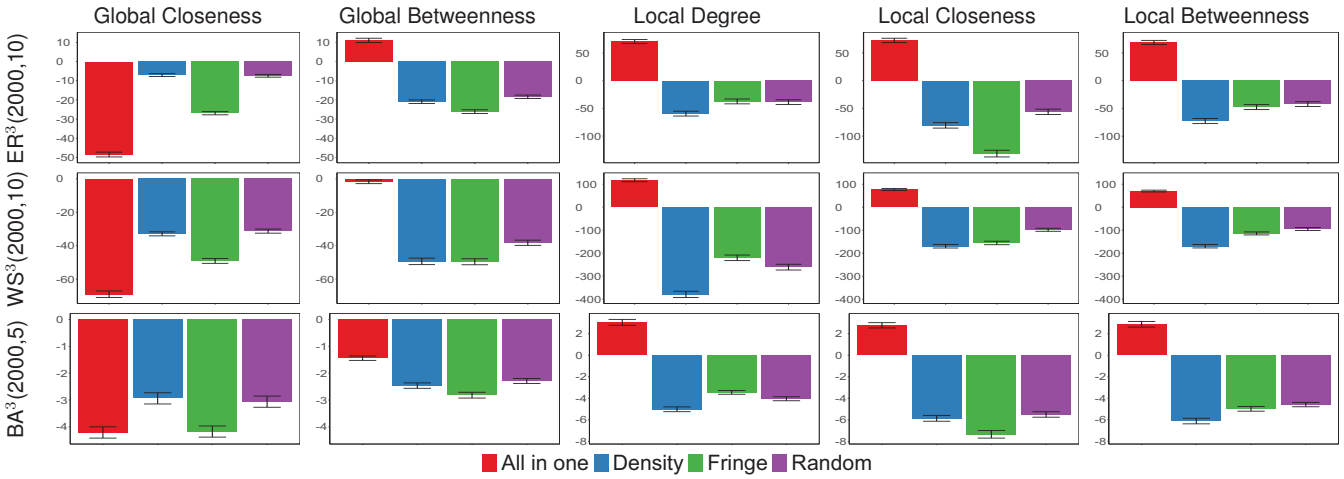

Figure 3: Given different centrality measures and different networks with 2000 nodes and 3 layers (ER—Erdős-Rényi, WS—Watts-Strogatz, BA—Barabási-Albert), the figure depicts the average change in centrality ranking of 10 different evaders as a result of execution of hiding heuristics. The experiment is repeated 100 times, with a new network generated each time. Error bars represent 95% confidence intervals. The results for the real-life networks can be found in Waniek et al. (2019).

Algorithm 1 "All in one" heuristic

Input: Multilayer network M, the evader $\widehat{v}$, contacts F
Output: Edges to be added to the network, i.e., the set A

$A \leftarrow \emptyset$
$F^* \leftarrow F$
$L^* \leftarrow \{\alpha \in L : \widehat{v} \in V^\alpha\}$
while $|F^*| > 0$ **do**
  $\alpha^* \leftarrow \arg\max_{\alpha \in L^*} |F^* \cap V^\alpha|$
  for $v \in F^* \cap V^\alpha$ **do**
    $A = A \cup \{(\widehat{v}^{\alpha^*}, v^{\alpha^*})\}$
  $F^* = F^* \setminus V^\alpha$
return A

Algorithm 2 Fringe heuristic

Input: Multilayer network M, the evader $\widehat{v}$, contacts F
Output: Edges to be added to the network, i.e., the set A

$A \leftarrow \emptyset$
$L^* \leftarrow \{\alpha \in L : \widehat{v} \in V^\alpha\}$
for $v \in F \cap V^\alpha$ **do**
  $\alpha^* \leftarrow \arg\min_{\alpha \in L^*} |N^\alpha(v) \setminus F|$
  $A = A \cup \{(\widehat{v}^{\alpha^*}, v^{\alpha^*})\}$
return A

be the neighbors of the evader in the original network. After that, we remove all original edges between the evader and those contacts, and act as if the evader was never connected to those individuals, but rather wants to connect to them while remaining hidden from centrality analysis. Finally, we connect the evader to the contacts using edges chosen by one of our heuristics. We record the difference between the ranking of the evader in the original, unchanged

Algorithm 3 Density heuristic

Input: Multilayer network M, the evader $\widehat{v}$, contacts F
Output: Edges to be added to the network, i.e., the set A

$A \leftarrow \emptyset$
$L^* \leftarrow \{\alpha \in L : \widehat{v} \in V^\alpha\}$
for $v \in F \cap V^\alpha$ **do**
  $\alpha^* \leftarrow \arg\max_{\alpha \in L^*} \frac{|\{w \in F : (\widehat{v}^\alpha, w^\alpha) \in A\} \cap N^\alpha(v)| + |F \cap N^\alpha(v)|}{\max(1, |\{w \in F : (\widehat{v}^\alpha, w^\alpha) \in A\}|)}$
  $A = A \cup \{(\widehat{v}^{\alpha^*}, v^{\alpha^*})\}$
return A

network, and in the network after running the heuristic. In so doing, we quantify the impact of strategically choosing the relationships to be formed with the group of contacts. Note that for the local centrality measures, we need to aggregate the centrality scores for each layer into a single ranking for the entire network. We do so by assigning to each node v the following centrality score: $\frac{1}{\min_{\alpha \in L} r^\alpha(v)}$, where $r^\alpha(v)$ is the ranking of v in layer α.

Simulation Results

The results of our simulations are presented in Figure 3 (see Waniek et al. (2019) for the description of the datasets). Each row corresponds to a network, and each column corresponds to centrality measure. Each bar represents the change in the evader's ranking after using a particular heuristic (the color of the bar corresponds to the heuristic being used). A negative change implies that the ranking of the evader decreased, i.e., she became more hidden. In contrast, a positive change implies that the heuristic backfired, i.e., the evader actually became more exposed.

As can be seen, there is no heuristic that dominates the others, i.e., no heuristic is superior against all centrality mea-

sures. The "All in one" heuristic proves to be effective in hiding from global closeness centrality in many cases. Unfortunately, if the network is analyzed with one of the local centrality measures, the evader may become even more exposed. For every considered centrality measure, either the Density or the Fringe heuristic is among the most effective methods for hiding, and they never make the evader more exposed. Finally, commenting on the results of the Random heuristic, they demonstrate that it is relatively effective to simply get rid of excess links (i.e., avoid connecting with each node in more than one layer) and spread the remaining connections uniformly.

Our results show also that the global centrality measures are on average much harder to hide from than their local counterparts. This demonstrates the importance of analyzing the entire structure of a multilayer network, rather than focusing on each layer separately.

Regarding the size of the networks used in the simulations, note that the heuristics use only local information and can be easily applied in much larger networks. However, the cost of computing complete rankings of the multilayer centrality measures, which is necessary for us to present our results, grows quickly with the size of the network. Hence, we present results for the networks of moderate size.

Conclusions

We studied the problem of evading centrality analysis in multilayer networks, and analyzed this problem both theoretically and empirically, thereby initiating the study of evading social network analysis tools in multilayer networks. Interesting future directions include developing more sophisticated heuristics for evading centrality measures, and analyzing the problem of evading link-prediction algorithms in multilayer networks.

Acknowledgments

Marcin Waniek was supported by the Polish National Science Centre grant 2015/17/N/ST6/03686. Tomasz Michalak was supported by the Polish National Science Centre grant 2016/23/B/ST6/03599.

References

Anthonisse, J. M. 1971. The rush in a graph. *Amsterdam: University of Amsterdam Mathematical Centre*.

Bavelas, A. 1948. A mathematical model for group structures. *Human organization* 7(3):16–30.

Beauchamp, M. A. 1965. An improved index of centrality. *Behavioral Science* 10(2):161–163.

De Domenico, M.; Solé-Ribalta, A.; Cozzo, E.; Kivelä, M.; Moreno, Y.; Porter, M. A.; Gómez, S.; and Arenas, A. 2013. Mathematical formulation of multilayer networks. *Phys. Rev. X* 3:041022.

EU. 2016. Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. *Official Journal of the European Union* L119:1–88.

Freeman, L. C. 1977. A set of measures of centrality based on betweenness. *Sociometry* 35–41.

Kayes, I., and Iamnitchi, A. 2015. A survey on privacy and security in online social networks. *arXiv preprint arXiv:1504.03342*.

Kearns, M.; Roth, A.; Wu, Z. S.; and Yaroslavtsev, G. 2016. Private algorithms for the protected in social network search. *Proceedings of the National Academy of Sciences* 201510612.

Kitchin, R. 2016. The ethics of smart cities and urban science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374(2083):20160115.

Kivelä, M.; Arenas, A.; Barthelemy, M.; Gleeson, J. P.; Moreno, Y.; and Porter, M. A. 2014. Multilayer networks. *Journal of complex networks* 2(3):203–271.

Lane, J. I.; Stodden, V.; Bender, S.; and Nissenbaum, H., eds. 2014. *Privacy, big data, and the public good: frameworks for engagement*. Cambridge University Press.

Michalak, T. P.; Rahwan, T.; and Wooldridge, M. 2017. Strategic social network analysis. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Mislove, A.; Viswanath, B.; Gummadi, K. P.; and Druschel, P. 2010. You are who you know: Inferring user profiles in online social networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, 251–260. New York, NY, USA: ACM.

Narayanan, A., and Shmatikov, V. 2009. De-anonymizing social networks. In *Security and Privacy, 2009 30th IEEE Symposium on*, 173–187. IEEE.

Persily, N. 2017. The 2016 us election: Can democracy survive the internet? *Journal of democracy* 28(2):63–76.

Shaw, M. E. 1954. Group structure and the behavior of individuals in small groups. *The Journal of Psychology* 38(1):139–149.

Waniek, M.; Michalak, T. P.; Rahwan, T.; and Wooldridge, M. 2017. On the construction of covert networks. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, 1341–1349. IFAAMAS.

Waniek, M.; Michalak, T. P.; Wooldridge, M. J.; and Rahwan, T. 2018. Hiding individuals and communities in a social network. *Nature Human Behaviour* 2(2):139.

Waniek, M.; Zhou, K.; Vorobeychik, Y.; Moro, E.; Michalak, T. P.; and Rahwan, T. 2019. How to hide one's relationships from link prediction algorithms. *Scientific Reports* 9.

Waniek, M.; Michalak, T. P.; and Rahwan, T. 2019. Hiding in multilayer networks. *arXiv* 1911.05947.

Yu, S.; Zhao, M.; Fu, C.; Huang, H.; Shu, X.; Xuan, Q.; and Chen, G. 2018. Target defense against link-prediction-based attacks via evolutionary perturbations. *arXiv preprint arXiv:1809.05912*.

Zhou, K.; Michalak, T. P.; Waniek, M.; Rahwan, T.; and Vorobeychik, Y. 2019. Attacking similarity-based link prediction in social networks. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, 305–313. IFAAMAS.

Zhou, B.; Pei, J.; and Luk, W. 2008. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM Sigkdd Explorations Newsletter* 10(2):12–22.