EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE

Open Access

Percolation framework reveals limits of privacy in conspiracy, dark web, and blockchain networks

Louis M. Shekhtman[1]*, Alon Sela[2] and Shlomo Havlin[3]

*Correspondence:
lsheks@gmail.com
[1] Network Science Institute, Northeastern University, Boston, USA
Full list of author information is available at the end of the article

Abstract

We consider the limits of privacy based on the knowledge of interactions in anonymous networks. In many anonymous networks, such as blockchain cryptocurrencies, dark web message boards, and other illicit networks, nodes are anonymous to outsiders, however the existence of a link between individuals is observable. For example, in blockchains, transactions between anonymous accounts are published openly. Here we consider what happens if one or more individuals in such a network are deanonymized by an outside investigator. These compromised individuals could then potentially leak information about others with whom they interacted, leading to a cascade of nodes' identities being revealed. We map this scenario to percolation and analyze its consequences on three real anonymous networks—(1) a blockchain transaction network, (2) interactions on the dark web, and (3) a political conspiracy network. We quantify, for different likelihoods of individuals possessing information on their neighbors, p, the fraction of accounts that can be identified in each network. We then estimate the minimum and most probable number of steps to a desired anonymous node, a measure of the effort to deanonymize that node. In all three networks, we find that it is possible to deanonymize a significant fraction of the network ($> 50\%$) within less than 5 steps for values of $p > 0.4$. We show how existing measures and approaches from percolation theory can help investigators quantify the chances of deanonymizing individuals, as well as how users can maintain privacy.

Keywords: Privacy; Percolation; Dark Web; Blockchain; Social networks

1 Introduction

For many forms of modern communication it is difficult for an individual to avoid participating in an open network where interactions can be observed by others [1–5]. This is also true for many networks where individuals carry out illicit interactions such as transactions through blockchain-based cryptocurrencies, communications on the dark web, and communications via anonymous email accounts or pre-paid cell phones. In all of these cases, the network itself is visible or can be obtained by authorities. Thus, the main method through which individuals, especially those carrying out illicit dealings, maintain anonymity is by ensuring that authorities cannot link them to their accounts in the anonymous observed network [6–8]. At the same time, in many of these networks the very act of interacting with another party via a network link will require the two individuals at the ends of the link to exchange some information that could enable them to identify each other [9]. Thus, if one individual is identified, it may, with some probability, be possible to obtain information on some of her neighbors with whom she interacted [10].

For example, if a physical item is purchased via an anonymous transaction network e.g., blockchain cryptocurrencies [11, 12], then the buyer must provide a shipping address to receive the item, sharing information that can reveal their location and therefore identity. Likewise, if two individuals communicate via phone or text messages, then it is likely that they possess some knowledge of the person on the other end of the line. Finally, in the case of online interactions, if user A's computer is hacked, then A's identity might become known to the hacker. The hacker could then search for additional information on A's computer e.g., email correspondence or private messages through online forums, to learn the identities of other individuals who interacted with A. After doing so, the hacker could then attempt to hack into the computers of these individuals (e.g., via a Trojan horse email) and potentially traverse the network of A's contacts. Furthermore, the very fact that A was hacked could also be useful in hacking the neighbors of A, as individuals are more likely to trust emails from a known source [13]. Similarly, in the context of criminal [14, 15], terrorist [16, 17] or conspiracy networks [18], if a set of individuals interact via anonymous communications, should one member of the group be identified, they could potentially be followed or interrogated leading to information on other individuals, who could then also be monitored to identify additional members of the network, and so on. Other motivating cases are users of burner phones who do not provide their names, but whose calls can be tracked; anonymous email accounts where a name is not provided, but a user's messages may be saved by the email provider; Telegram messenger, where anonymous users interact in public groups; and other criminal or conspiracy networks.

Some approaches have considered deanonymizing the individuals behind nodes in a network, specifically in the context of cryptocurrencies [19]. For example, [20] identified a few heuristics that can link multiple accounts as belonging to the same individual and others have noted that by studying the time when a transaction is submitted to the blockchain network an account's identity and IP address can be determined [21]. However, users have often found ways to overcome these issues, such as by using private browsers like Tor to access the bitcoin network [22]. Outside of cryptocurrencies, several works have considered how to identify individuals based on their interactions [9, 23–26], though these works have not presented a quantitative framework to assess the chances of successful identification, the fraction of nodes that can be identified, and the effort necessary to do so. Here we show that the question of anonymity of network actors, and the corresponding ability of a party seeking to deanonymize the individuals based on information from their neighbors, can be analyzed and quantified using tools and methods from *percolation theory* of statistical physics [27–30]. Furthermore, we demonstrate that classical quantities from percolation theory provide crucial methods to quantify the extent to which anonymity can be maintained among individuals in real networks. Aside from the giant component, we also explore path lengths as a proxy of the effort and resources needed to identify an individual, as well as consider a search process to estimate a 'realistic' path length that incorporates the fact that investigators are likely to reach dead ends along the way in their

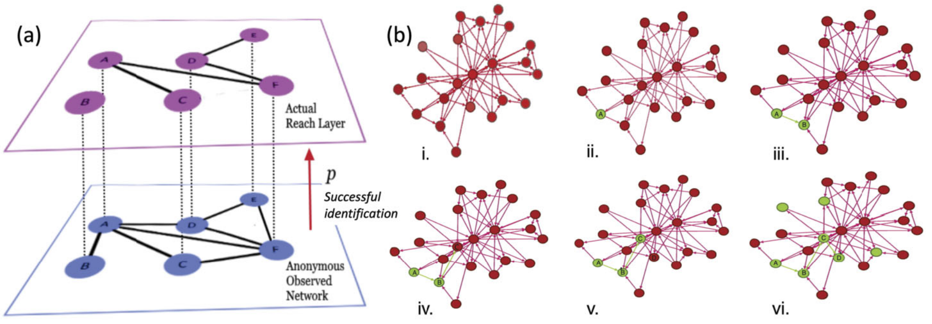

Figure 1 *Giant Component and deanonymization.* **(a)** Illustration of an anonymous network where the observed layer (blue) represents the network of interactions between anonymous actors; while the deanonmyizable layer (pink) is the subset of interactions where individuals have knowledge about their neighbors that an investigator could exploit. Theoretically, based on the *observed* network an investigator who knows the identity of node A could use this information to obtain the identity of node E by first determining the identity of one intermediate node (D). However, in practice, based on the deanonymizable layer the investigator would need two intermediate nodes (F and D) to identify node E since D did not reveal their identity to A in their interaction. **(b)** Demonstration of the process through which deanonymization could occur. In the initial network (**i**) no users are known. In (**ii**) after investigators successfully identify node *A*, this node is marked green meaning it is known and investigators can then use information from node *A* to try and traverse the network. Initially all links are unknown and thus marked red. Green links represent links where users have provided information about the neighbor at the other end and green nodes are identified nodes. In (**iii**) node *B* is successfully identified through information obtained via a link from node *A*. Next, in (**iv**) node *B* provides information about node *C*, who is also shown as identified in subplot (**v**). Finally, node *C* provides information leading to the identification of node *D* in (**vi**)

investigation. In contrast, to previous works, our approach is fundamental to the nature of privacy interactions when counter-parties must have information about one another and the network of interactions can be observed.

We demonstrate our general framework in Fig. 1, enabling us to quantify the extent to which information on the network can be exploited and the likelihood of individuals being identified by their neighbors. We apply our percolation theory approach on three examples of real-world anonymous networks: (i) a network of transactions from a blockchain-based cryptocurrency [31], (ii) a network of interactions related to illegal activities via the dark web [32], and (iii) a political conspiracy network [18]. In all three of these networks, the question of anonymity is very important: cryptocurrencies and the dark-web are often used by criminal organizations [20, 33–35], whereas conspirators depend on not being uncovered in order to avoid criminal charges.

2 Theoretical framework

As explained above, we seek to understand how investigators could leverage identifying information exchanged between interacting parties to uncover specific individuals. Nonetheless, in some cases, an individual may not have any identifiable information about the party with whom they interacted via a link in the network. In this sense, one could consider such a link 'failed' in the sense that no deanonymization can be carried out via that link. Thus, the probability of links exchanging identifying information can be mapped to the link-occupation probability p from percolation theory [36–40]. In percolation theory, a key quantity of interest is the fractional size of the largest connected component (giant component) S as a function of p. In the context of anonymity, the giant component represents the set of nodes that could all be revealed (given sufficient time and effort) if one of them is discovered. Furthermore, the value of p where S grows to a macroscopic size, typically referred to as the critical point p_c [38], can serve as a simple measure for estimating whether a network is likely to allow for deanonymization or not.

Similarly, for smaller connected components, each component is a set of nodes where if any one of them is deanonymized, then the rest of the set could (with sufficient effort) also be identified. Therefore the total number of components, n_{comp}, represents the minimal number of source nodes (in distinct components) that need to be identified independently in order to deanonymize the entire network. Similarly if one has a set of source nodes $n_{1,2,\ldots,k}$ each in a different component, then the total number of nodes that can be identified is the sum of all k separate components $T_S = |S_1| + |S_2| + \cdots + |S_k|$, where T_S is the total number of nodes that can be identified (see Fig. S2). As we show, in our datasets at a practical level, the small components tend to be of insignificant size compared to the giant component for all values of $p > 0$ (Fig. S3). For example, the giant component S captures approximately 99.9% of the nodes for the Blockchain network, 100% of the nodes for the Dark Web network, and 76.5% of the nodes for the Conspiracy network. Also the small components are only connected to a limited number of other nodes, and thus are less valuable for investigators.

A key aspect of deanonymization is the effort required to deanonymize a particular node, which is reflected in the actual number of interrogations that an investigator must carry out to deanonymize a target node, given the identity of some source node or nodes. For example, it is possible that investigators might identify one dark web user posting large amounts of illicit content or a specific account participating in suspicious transactions on the Blockchain. The investigators could begin by seeking information from an identified source node about her neighbors and then moving on to the neighbors' neighbors and so on along the shortest path, until reaching the desired node of interest. The minimum number of individuals that must be identified to reach the desired node is thus the shortest path length, l, from the source node(s) to the target node. This reflects a measure of effort since each individual that must be identified along the path will require dedication of resources for monitoring, questioning, etc.

Lastly, we generalize the above shortest path measure to include the fact that some individuals along the shortest path will not have or supply identifying information on all of their neighbors. Therefore, we propose a greedy algorithm, described later, where the investigators first interrogate the nodes along the initial shortest path from their source node(s) to the target and then update to the next (new) shortest path if the investigation reaches a dead end i.e., a link where the node on the other end cannot be identified. We define the number of steps along the paths using this greedy algorithm as ℓ_{actual} as it approximates what could be a possible 'actual' number of inquiries needed to reach a specific node given that investigators do not know in advance which links are useful and which will ultimately lead to a dead end. These and other relevant measures from percolation theory are described in Table 1.

3 Results

We apply the framework described above on three publicly available anonymous networks: (i) the flow of funds within the Ethereum blockchain-based cryptocurrency [31], which is the second largest cryptocurrency by market cap; (ii) A forum of users participating in sharing of child pornography on the dark web [32]; and (iii) A network of political

Table 1 Mapping to percolation. We demonstrate and provide brief explanations on the various measures used and how they relate to the traditional measures from percolation theory

Parameter	Percolation Theory Definition	Privacy Interpretation
p	Fraction of occupied links	Probability a node can identify its neighbors
S	Giant connected component (GCC)	Largest group of mutually vulnerable nodes
SG	Second largest component	Second largest group of vulnerable nodes
p_c	Critical point	p for which deanonymization is feasible
n_{comp}	Num. Components	Min. num. of sources to identify whole network
l	Shortest Path Length	Optimal number of interrogations
ℓ_{actual}	Greedy Algorithm Path Length	Realistic number of interrogations

Table 2 Data used and basic real-world network characteristics. The Blockchain network is a network of transactions between cryptocurrency accounts in Ethereum, the Dark Web network is a network of users sharing child pornography on the dark web, and the Conspiracy network is a network of political conspirators in Brazil. N is the number of nodes, V is the number of links, $\langle k\rangle$ is the average degree, c is the clustering coefficient, $\langle \ell\rangle$ is the average shortest path length (which for the Blockchain network could only be estimated due to the network size), and density is the overall density of links defined by the number of links in the network divided by the number of all possible links. *Median is listed rather than mean due to fat-tail

Metric/Network	Blockchain	Dark Web	Conspiracy
N	2,291,941	10,407	404
V	5,262,468	820,272	3350
$\langle k\rangle$	2.8	150	9.7
c	0.21	0.83	0.85
$\langle l\rangle$	4*	2.15	2.98
Density	10^{-6}	0.0076	0.022

conspirators in Brazil [18]. A summary of basic statistics on these datasets is available in Table 2.

Using the proposed percolation approach we analyze and quantify the sizes of the largest component in the three real-world networks as a function of p, the likelihood that a node can identify its neighbor. This largest component corresponds to the mean fraction of accounts that can be deanonymized, as a function of p, after a single source node is identified.

In Fig. 2(a)–(c), we show the fractional size of the giant connected component, S, and the second largest connected component, SG, for each of the three networks. We see that the largest component S typically constitutes a large fraction of the network, suggesting that most individuals in the network can be identified via information from others. For example in the Dark Web and Conspiracy networks, even for p values near 0.1, 90% of individuals are in the giant component and can be identified. For the Blockchain network this fraction is lower, but even for $p = 0.5$ around 75% of accounts are identifiable in the giant component. The fact that only near $p \to 0$ does $S \to 0$ is typical for networks with long-tailed or scale-free degree distributions containing hubs, which is true for the degree distribution of all 3 of the networks shown here (see Fig. S1) [41–43]. Furthermore, the second largest component is typically quite small (max of 0.07 for the Conspiracy network and smaller for the others).

It is worth noting that in many networks, the likelihood of uncovering links will not be homogeneous, and rather that some links are more likely to involve exchanging identifying information than others. Such effects are already incorporated in our subsequent analysis of the Blockchain network, where exchanges are assumed to have more knowledge of their

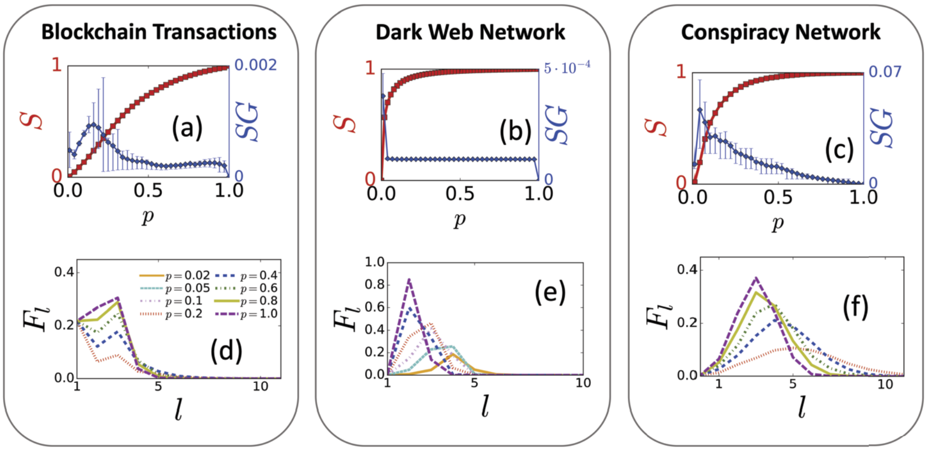

Figure 2 *Fraction of Reachable Nodes and Minimum Effort to reach them.* (**a**)–(**c**) The fractional size of the largest, *S*, and second largest, *SG*, components in each of our three real networks. Points are averages of 100 realizations and error bars represent one standard deviation. For *S* the error bars are smaller than the size of the points. (**d**)–(**f**) The fraction of nodes, F_l, reached at l steps (shortest path) for each of our networks. For the Blockchain Transaction Network we assume that a set of 10 nodes belonging to 5 exchange companies are sources, whereas for the Dark Web Network and the Conspiracy Network, we randomly select individual source nodes. For all 3 of our networks, the shortest paths are very short, with most reachable nodes, being identified within $l \approx 3$ or ≈ 4 steps. Values of $p \leq 0.1$ are shown only for the Dark Web because online interactions are less likely to involve identifying information

customers than other nodes. However, it is also worth exploring heterogeneity in the context of the Dark Web network. Prior work on this network [32] found that there are two groups of nodes, with a small subset of nodes acting as a 'core' that shares illicit content and a larger set of nodes acting only as consumers. We suggest that a core node is more likely to have information on a regular node, whereas a regular node will have information on a core node with a much lower probability, as core nodes are likely more careful about who they share information with i.e., $p_{\mathrm{core}\rightarrow\mathrm{reg}} > p_{\mathrm{reg}\rightarrow\mathrm{core}}$. In Fig. 3 we show the size of the giant component as we vary both of these likelihoods. We find that as both likelihoods increase up to around 10% there is a considerable increase in the fraction of the network that can be identified and that when $p_{\mathrm{core}\rightarrow\mathrm{reg}}$ reaches around 20% the fraction of nodes that can be identified starts to plateau.

3.1 Results—shortest paths

We next consider the shortest path lengths, a proxy for the effort to deanonymize an individual. For the Blockchain Transaction Network, rather than considering shortest paths between a randomly selected source and a target, we use a set of 10 nodes belonging to 5 different so-called 'exchanges,' that convert cryptocurrency to fiat currency, as our source nodes. This is because exchanges inherently know the identities of their neighbors due to legal policies they have in order to prevent money laundering (know-your-customer policies) [44], and moreover, they are the hubs of the Blockchain Transaction Network making them worthwhile targets for investigators [31] (for more on this choice, see Additional file 1). In Fig. 2(d)–(f) we show the fraction of nodes found at l steps, S_l, for different values of p, for each of the three analysed networks. For all 3 networks, the ideal shortest paths are very short (suggesting they possess the 'small-world' property [45]), with most reachable nodes, being reached within $l \approx 3$ or 4 steps. For the Dark Web network we explore

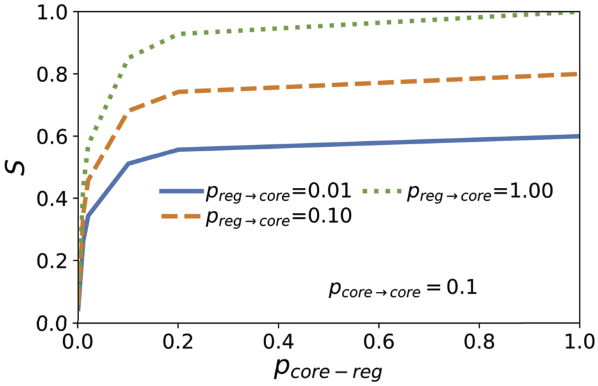

Figure 3 *Heterogeneous probability of uncovering links.* We show the size of the giant component for heterogeneous likelihoods of nodes knowing their neighbors in the Dark Web network. We segregate nodes into those who are part of the network 'core' and that share illicit content versus the remainder of nodes that do not share content. We then consider different likelihoods of nodes having information on their neighbors based on whether the link is from a core node to a non-core node $p_{\text{core}\to\text{reg}}$, from a non-core node to a core node $p_{\text{reg}\to\text{core}}$, or between two core nodes $p_{\text{core}\to\text{core}}$. We show the giant component, S as a function of the likelihood that core nodes can identify regular node neighbors and consider different likelihoods that a regular node can identify a core node. We find a rapid increase in the fraction of nodes that can be identified as both $p_{\text{reg}\to\text{core}}$ and $p_{\text{core}\to\text{reg}}$ reach around 10%. We chose a fixed value of the likelihood that core nodes can identify other core nodes of 10%. Results are averages of 1800 simulations on the Dark Web network

several smaller values of p because interactions in the Dark Web are less likely to involve individuals having information on one another compared to the case of transferring funds or cooperating in a conspiracy. These short paths, suggest that deanonymizing individuals based on their neighbors is feasible in many cases and would not require burdensome levels of investigator resources. It is worth mentioning that given only the observed network structure, investigators can determine if the network is a small-world based solely on the structure without carrying out deanonymization efforts, potentially helping to determine whether such efforts are worthwhile.

3.1.1 Analytic approach to shortest paths

Investigators can further leverage an analytic approach, presented next, to estimate the fraction of nodes that appear to have some given path length ℓ from the source to determine how likely the node is to actually be reached in ℓ steps. First, we note that one can naively estimate the likelihood that the path between the source node and some target node will continue to exist after $1 - p$ fraction of links are removed. This is simply,

$$P_\ell(p) = p^\ell, \tag{1}$$

where $P_\ell(p)$ is the likelihood that the path of length ℓ continues to exist, p is the fraction of links that involved identifying information, and ℓ is the length of the path. This statement simply says that the path exists if and only if every link along it exists. This initial estimate serves as a lower bound on the likelihood that a path of length ℓ exists between the source and target since there can be other fully or partially non-overlapping paths that are also of length ℓ. In Fig. 4(a) we see that for the Conspiracy network the difference between the theory of Eq. (1) is fairly small, while in Fig. 4(b) we see that for the Dark Web Network there is a larger divergence. It is worth mentioning that for $\ell = 1$ the theory is exact in

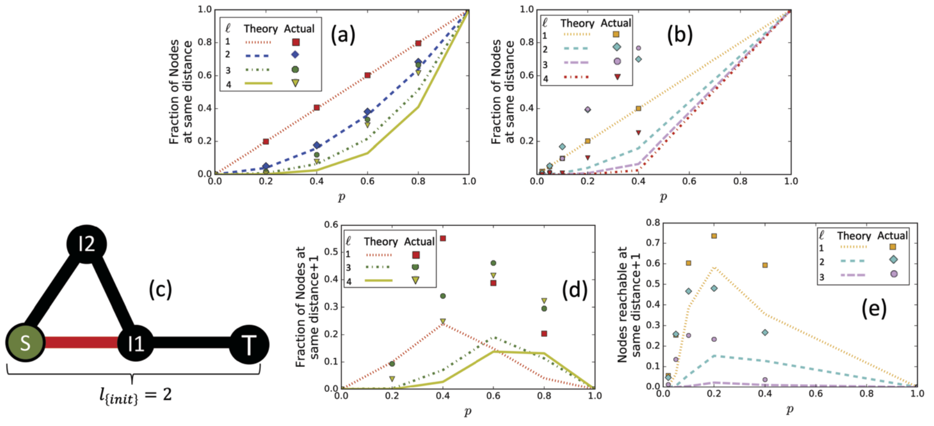

Figure 4 *Theoretical approach to path lengths.* (**a**) We show our theory for of what fraction of nodes having an initial path length of ℓ will remain at that path length if only p fraction of links involve identifying information in the Conspiracy Network. (**b**) Same for the Dark Web Network. (**c**) We consider the case where one link along the original path fails (from $S \rightarrow I_1$), but there is a simple 'detour' via I_2 leading to an increase of 1 in the path length. (**d**) We show our theory what fraction of nodes will be at the original path length plus one (based on the 'detour' suggested in c) for the Conspiracy network. (**e**) Same as (**d**) but now for the Dark Web Network

all cases since the direct link between the two nodes either exists or doesn't exist with probability p.

We next consider a simple case that can be addressed analytically where only a single link on the path fails, yet a 'detour' going around that link via another intermediate node exists, see Fig. 4(c). To give a naive estimate for the likelihood of this case arising, we first estimate the likelihood that only a single link on a path of length ℓ fails, which is $\ell p^{\ell-1}(1-p)$ using the binomial expansion. If we assume that the node at the end of the failed link has $\langle k \rangle$ links, then we need only know the likelihood that one of the $\langle k \rangle - 1$ other links leads to a node who connects back to the original path. This likelihood is given by the clustering coefficient c, thus we can estimate that there are $(\langle k \rangle - 1) \cdot c$ paths that circumvent the failed link. Each of these paths are of length 2 and so the likelihood that one survives is p^2. We now want to know the likelihood that at least one of these paths exists, which is given by one minus the likelihood that none of them exist, or $1 - (1 - p^2)^{c(\langle k \rangle - 1)}$. Incorporating all of the terms leads to

$$P_{\ell+1}(p) = \ell p^{\ell-1}(1-p)\left[1 - \left(1 - p^2\right)^{c(\langle k \rangle - 1)}\right], \tag{2}$$

where $P_{\ell+1}(p)$ is the likelihood that the actual shortest path is one greater than the initial shortest path with $p = 1$. In Fig. 4(d)–(e) we show the theory and actual calculations for the Conspiracy and Dark Web Networks, finding that while Eq. (2) significantly underestimates the likelihood that the path length will increase by one (since there are many other ways to arrive at a path length of $\ell + 1$), it does preserve the same shape of the curve as the actual results.

3.2 Results—realistic paths

However, the shortest path length is only the *minimal* number of interrogations an investigator would need if they had perfect knowledge in advance about which individuals have information on which neighbors. In practice, investigators will reach individuals who do not possess identifying information on their neighbors or do not reveal information they possess. Therefore, we propose a greedy algorithm that investigators could use in order to carry out their investigation efficiently and estimate the effort required to deanonymize a particular individual. Essentially, our greedy algorithm begins by investigating along the initial shortest path between the source node(s) and the target. The investigators then interrogate the nodes along the shortest path until they hit a dead end i.e., reach a node that does not provide identifying information about the desired neighbor. They then remove that link from the network, calculate the new shortest path from the source to the target node in the modified network, and attempt to traverse the new shortest path. This process can be done iteratively until the target node is reached or until it is determined that no other possible paths exist.

We formally write out our greedy algorithm in Additional file 1, Algorithm S1, and in Fig. 5(a)–(c) we demonstrate the process of our algorithm visually. If all of the links on the original shortest path are indeed identifiable then our algorithm will lead to the minimal number of interrogations. In contrast, if we hit dead ends along the various paths that we pursue, then our algorithm will lead to a greater number of interrogations than in the idealized case where investigators have complete information.

We apply the proposed greedy algorithm to the three datasets. As before, for the Blockchain Transaction Network, we consider 10 nodes belonging to 5 exchanges as our source nodes (with all of their links known), whereas for the Dark Web Network and Conspiracy Network we choose a single source node randomly. We then choose random target nodes and assess how many actual steps are required to reach the target for different values of p. To understand the network effects, we focus on target nodes that are at least $l \geq 2$ from the source i.e, not direct neighbors of the source node.

In Fig. 5(d)–(f) we show the fraction of nodes, F_A, reached for a given number of steps ℓ_{actual}. We find that the investigators' lack of knowledge about which links can reveal identifying information on a neighbor can be a significant detriment to their ability to optimally traverse the network for low values of p. This is observed from the fact that the distribution shifts significantly to the right (higher values of ℓ_{actual}) compared to the optimal case in Fig. 2(d)–(f). For higher levels of p, the detriment is less pronounced as most of the links involve identifiable information and the original observed shortest path is likely to be optimal.

Once an investigator has reached k individuals and not identified the target, they face a sunk-cost problem as they do not know how close, if at all, they are to identifying the target. To assess this situation, we considered the likelihood that after interrogating k individuals, continuing to investigate will lead to the identification of the desired target. In Fig. 5(g)–(i) we show this likelihood, P_{Find}, for differing levels of p on each of our networks. We see that in all three networks, the likelihood of ultimate success tends to decrease as ℓ_{actual} increases, suggesting that it would make sense for investigators to place a limit on how many inquiries they will carry out. For example, in the Dark web network for $p = 0.2$, for $\ell_{\text{actual}} = 4$ there is still an over 90% chance of successfully identifying a node when continuing, yet for $\ell_{\text{actual}} = 15$ there is only a 70% of identifying the node. Results for a scale-free network model are qualitatively similar to those in our real-networks suggesting that hubs play an important role in the observed results (Fig. 5(j)–(k)). We further see that when we compare Erdős-Rényi and scale-free networks with $\gamma = 2.5$ and $\gamma = 3$, that the distribution of F_A and P_{Find} changes considerably with Erdős-Rényi networks tending to

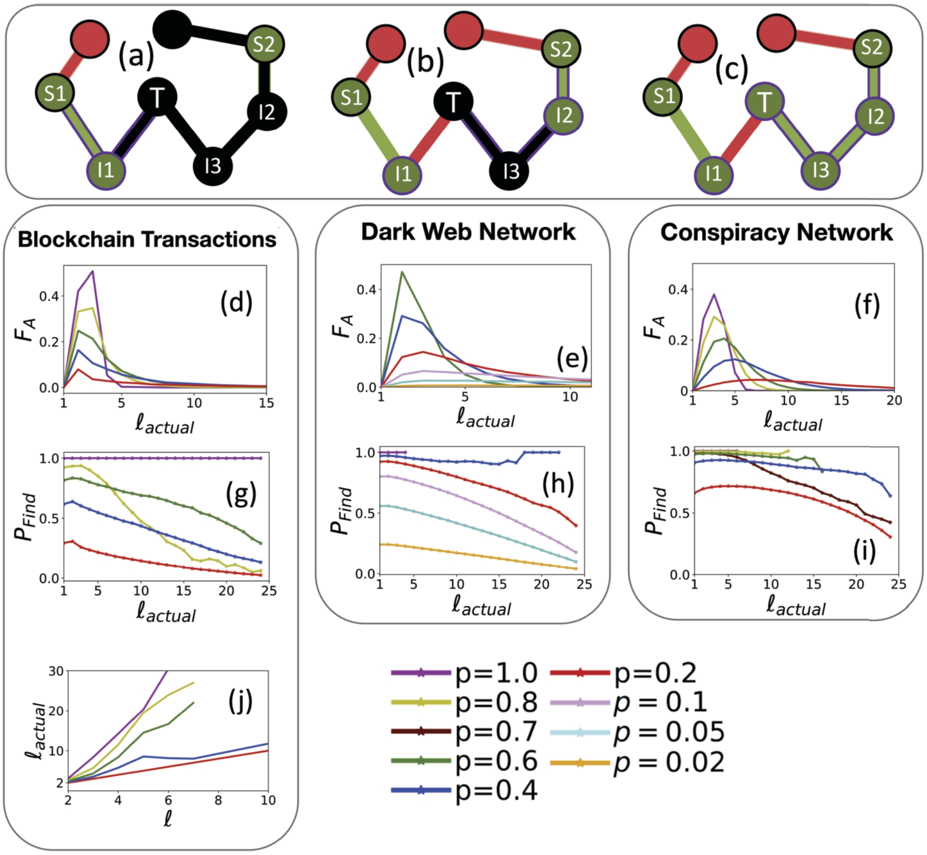

Figure 5 *Realistic effort to identify an anonymous node.* (**a**)–(**c**) Illustration of the proposed greedy algorithm for deanonymizing a target node T, given identities of source nodes $S1$ and $S2$. In (**a**) we find the shortest path, $S_1 \rightarrow I_1 \rightarrow T$ (links of the shortest path outlined in purple) and interrogate the first node I_1. (**b**) After S_1 identifies I_1, we reach a dead end as node I_1 does not provide information on T. We then again calculate the shortest path to T and find a path $S_2 \rightarrow I_2 \rightarrow I_3 \rightarrow T$ (with its links highlighted). (**c**) We interrogate I_2 and then I_3 (outlined in purple again), who provide information leading to the target. While the optimal path would have taken only $l = 3$ steps, the actual path uses $\ell_{\text{actual}} = 4$ steps due to the extra step from $S_1 \rightarrow I_1$. (**d**)–(**f**) The fraction of nodes reached, F_A, for each value of ℓ_{actual} using our algorithm. Results are from at least 20,000 simulated source-target pairs (for different values of p). (**g**)–(**i**) For all three networks the likelihood that continuing to search after ℓ_{actual} steps will lead to successful identification of a given target node, P_{Find}. Note that less than 100 nodes have $\ell_{\text{actual}} > 15$ leading to high fluctuations. (**j**) The mean value of ℓ_{actual} as a function of the optimal value, ℓ, for the Blockchain Transactions. For lower levels of p, ℓ_{actual} tends to increase faster with ℓ. In all panels, only target nodes that are not direct neighbors of the source node(s) are considered

have considerably lower values of P_{Find} than a scale-free network, see Fig. S6. Furthermore, even increasing the exponent from $\gamma = 2.5$ to $\gamma = 3$ leads to fairly significant changes. Given that the Dark Web and Blockchain networks have values of $\gamma < 3$ it makes sense that our results are most similar to the case of $\gamma = 2.5$. For the Conspiracy network, there are too few nodes to observe a clear scaling in the tail, however there are clearly some hubs with degrees much larger than other nodes.

4 Discussion

Our mapping of deanonymization to percolation reveals that hubs, which exist in all three of the anonymous networks, play an important role in enhancing deanonymization. Hubs are common in many networks and they exacerbates privacy issues since they can potentially be identified via their many spokes and can then reveal information on the remainder of the network. It is also likely that hubs are common in many illicit transactions on the Dark Web and other criminal networks since only a few individuals are involved in core criminal activity whereas most people are likely only marginally involved. For example, in the Dark Web Network based on a child pornography ring, most of the nodes are presumably consumers of such content whereas the hubs are distributors with deeper criminal involvement. The issue of hubs is also significant for the Blockchain Transaction network where the hubs are exchanges which collect information on their neighbors due to know-your-customer policies.

A particularly unique property of the Blockchain network is that very long chains exist in it of accounts involved in only single transactions with others [20]. This is seen in that while the average path length in the Blockchain network is 19.3, its median path length is only 4; and while the 95% percentile of path lengths is only 6, the 99% percentile jumps to a path length of 474. These long chains were once suggested as a method of potentially obscuring an individual's identify by making it harder to track them [20]. However, in the setting we describe, such links are unlikely to actually improve privacy as they are easy to identify and only artificially lengthen the shortest path while not actually increasing the number of parties between the target node and source node since questioning the same individual would identify many of the intermediate nodes. In fact, these long chains could be a signal of suspicious activity and the beginning or ends of such chains could serve as target nodes worth trying to reach.

Some limitations should be noted. First, while estimating p for crypto network is difficult, we provide some proxies from other domains where computer or other illicit networks become comprised. These are described in the Additional file 1 section for areas regarding hacking success rates, click-through rates for marketing materials (which could be used to embed trojans), and other areas. Whether these estimates are transferable to our areas is still unknown, however they do provide some context on possible values of p and suggest estimates of p from 1–10%. A further limitation arises within the context of cryptocurrencies, where newer cryptocurrencies like Monero (https://getmonero.org) might be able to obscure the nature of the network by adding false transactions to the list of transactions, however in many cases identifying many real transactions is still possible with simple heuristics [46]. To apply our percolation framework one could then reduce the network to the known transactions, giving investigators at least some picture of which users might be identifiable. Furthermore, as individuals are uncovered, the counterparties to their transactions will become known, providing additional knowledge of the network structure and how it can be traversed.

Our work has demonstrated the feasibility of using information from particular sources to identify their associates in multiple anonymous networks. A framework similar to ours could be applied at the outset of an investigation to predict the likely resources necessary (number of interrogations/intermediate parties to be followed) in order to identify a particular anonymous actor in the network. Furthermore, our framework could be applied to other contexts like terrorist networks and intercepted communications from burner phones where the individuals behind those numbers are not known. Likewise, it enables ordinary users to assess their level of anonymity and highlights the importance of users maintaining anonymity even when interacting with a trusted party.

Further work could expand our analytic theory in Fig. 4(d)–(e) to also consider the degree specific clustering coefficients and average over the degrees and their prevalence to obtain a better estimate of how many nodes are reachable at distance $\ell + 1$. Likewise, future work could improve upon our greedy algorithm, which is an upper bound on the amount of effort needed. In particular, it does not use any metadata that may be associated with the links and incorporating such metadata may suggest that paths other than the shortest path should be pursued. For example, one could look create a scoring algorithm for links that includes the frequency that the different links appear e.g. how often transactions are made or how often two individuals communicate. Incorporating such information could lead to lower values of ℓ_{actual} than currently found in our algorithm. Finally, our work assumes a simplified model where only one piece of information is needed from a single other user to deanonymize an individual, however in some contexts information from multiple users could be combined to identify an individual. This can be compared to color-avoiding percolation where information was sent along different paths to avoid detection [47–49] and the spreading of complex contagions [50] where multiple nodes have to be activated to spread to a neighbor leading to complex cascades [51] of identification.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-023-00392-8.

Additional file 1. Supplementary information (PDF 5.5 MB)

Acknowledgements
Not applicable.

Funding
This work was supported by the BIU Center for Research in Applied Cryptography and Cyber Security in conjunction with the Israel National Cyber Directorate in the Prime Minister's Office and the Ariel Cyber Innovation Center in conjunction with the Israel National Cyber directorate in the Prime Minister's Office. S.H. also thanks the Israel Science Foundation, (Grant No. 189/19) and the joint China-Israel Science Foundation (Grant No. 3132/19), ONR, BSF-NSF, and DTRA (Grant No. HDTRA-1-19-1-0016) for support.

Abbreviations
Not applicable.

Availability of data and materials
Data is available from the original sources in [18, 31, 32]. Code for carrying out our simulations is available at: https://github.com/lsheks/percolation-deanonymization.

Declarations

Competing interests
The authors declare no competing interests.

Author contributions
All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by LMS; writing LMS, AS and SH. All authors read and approved the final manuscript.

Author details
[1]Network Science Institute, Northeastern University, Boston, USA. [2]Department of Industrial Engineering, Ariel University, Ariel, Israel. [3]Department of Physics, Bar-Ilan University, Ramat Gan, Israel.

Received: 2 May 2022 Accepted: 15 May 2023 Published online: 29 May 2023

References

1. Zheleva E, Getoor L (2009) To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In: Proceedings of the 18th international conference on world wide web. ACM, New York, pp 531–540

2. Bagrow JP, Liu X, Mitchell L (2019) Information flow reveals prediction limits in online social activity. Nat. Hum. Behav. 3(2):122
3. Allard A, Hébert-Dufresne L, Young J-G, Dubé LJ (2014) Coexistence of phases and the observability of random graphs. Phys. Rev. E 89(2):022801
4. Yang Y, Wang J, Motter AE (2012) Network observability transitions. Phys. Rev. Lett. 109(25):258701
5. Onnela J-P, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész J, Barabási A-L (2007) Structure and tie strengths in mobile communication networks. Proc. Natl. Acad. Sci. 104(18):7332–7336
6. Gross R, Acquisti A (2005) Information revelation and privacy in online social networks. In: Proceedings of the 2005 ACM workshop on privacy in the electronic society, pp 71–80
7. Garcia D (2017) Leaking privacy and shadow profiles in online social networks. Sci. Adv. 3(8):e1701172
8. Fergal R, Harrigan M (2013) An analysis of anonymity in the bitcoin system. In: Security and privacy in social networks. Springer, Berlin, pp 197–223
9. Xu JJ, Chen H (2004) Fighting organized crimes: using shortest-path algorithms to identify associations in criminal networks. Decis. Support Syst. 38(3):473–487
10. Barucca P, Caldarelli G, Squartini T (2018) Tackling information asymmetry in networks: a new entropy-based ranking index. J. Stat. Phys. 173(3–4):1028–1044
11. Böhme R, Christin N, Edelman B, Bitcoin TM (2015) Economics, technology, and governance. J. Econ. Perspect. 29(2):213–238
12. Pappalardo G, Di Matteo T, Caldarelli G, Aste T (2018) Blockchain inefficiency in the bitcoin peers network. EPJ Data Sci. 7(1):30
13. Moody GD, Galletta DF, Dunn BK (2017) Which phish get caught? An exploratory study of individuals' susceptibility to phishing. Eur. J. Inf. Syst. 26(6):564–584
14. Duijn PAC, Kashirin V, Sloot PMA (2014) The relative ineffectiveness of criminal network disruption. Sci. Rep. 4:4238
15. Toth N, Gulyás L, Legendi RO, Duijn P, Sloot PMA, Kampis G (2013) The importance of centralities in dark network value chains. Eur. Phys. J. Spec. Top. 222(6):1413–1439
16. Krebs VE (2002) Mapping networks of terrorist cells. Connections 24(3):43–52
17. Carley KM, Lee J-S, Krackhardt D (2002) Destabilizing networks. Connections 24(3):79–92
18. Ribeiro HV, Alves LGA, Martins AF, Lenzi EK, Perc M (2018) The dynamical structure of political corruption networks. J. Complex Netw. 6(6):989–1003
19. Henry R, Herzberg A, Kate A (2018) Blockchain access privacy: challenges and directions. IEEE Secur. Priv. 16(4):38–45
20. Ron D, Shamir A (2013) Quantitative analysis of the full bitcoin transaction graph. In: International conference on financial cryptography and data security. Springer, Berlin, pp 6–24
21. Koshy P, Koshy D, McDaniel P (2014) An analysis of anonymity in bitcoin using p2p network traffic. In: International conference on financial cryptography and data security. Springer, Berlin, pp 469–485
22. Das D, Meiser S, Mohammadi E, Kate A (2018) Anonymity trilemma: strong anonymity, low bandwidth overhead, low latency-choose two. In: 2018 IEEE symposium on security and privacy (SP). IEEE Press, New York, pp 108–126
23. Lovato J, Allard A, Harp R, Hébert-Dufresne L (2020) Distributed consent and its impact on privacy and observability in social networks. arXiv preprint. arXiv:2006.16140
24. Sarvari H, Abozinadah E, Mbaziira A, McCoy D (2014) Constructing and analyzing criminal networks. In: 2014 IEEE security and privacy workshops. IEEE Press, New York, pp 84–91
25. Malm A, Bichler G (2011) Networks of collaborating criminals: assessing the structural vulnerability of drug markets. J. Res. Crime Delinq. 48(2):271–297
26. Sparrow MK (1991) The application of network analysis to criminal intelligence: an assessment of the prospects. Soc. Netw. 13(3):251–274
27. Newman M (2010) Networks: an introduction. Oxford University Press, London
28. Cohen R, Havlin S (2010) Complex networks: structure, robustness and function. Cambridge University Press, Cambridge
29. Castellano C, Fortunato S, Loreto V (2009) Statistical physics of social dynamics. Rev. Mod. Phys. 81(2):591
30. Barabási A-L et al (2016) Network science. Cambridge University Press, Cambridge
31. Chen T, Zhu Y, Li Z, Chen J, Li X, Luo X, Lin X, Zhange X (2018) Understanding Ethereum via graph analysis. In: IEEE INFOCOM 2018-IEEE conference on computer communications. IEEE Press, New York, pp 1484–1492
32. Requião da Cunha B, MacCarron P, Passold JF, dos Santos LW, Oliveira KA, Gleeson JP (2020) Assessing police topological efficiency in a major sting operation on the dark web. Sci. Rep. 10(1):1–10
33. Ober M, Katzenbeisser S, Hamacher K (2013) Structure and anonymity of the bitcoin transaction graph. Future Internet 5(2):237–250
34. Meiklejohn S, Pomarole M, Jordan G, Levchenko K, McCoy D, Voelker GM, Savage S (2013) A fistful of bitcoins: characterizing payments among men with no names. In: Proceedings of the 2013 conference on Internet measurement conference. ACM, New York, pp 127–140
35. De Domenico M, Arenas A (2017) Modeling structure and resilience of the dark network. Phys. Rev. E 95(2):022313
36. Stauffer D, Aharony A (2018) Introduction to percolation theory. CRC Press, Boca Raton
37. Stanley HE (1973) Introduction to phase transitions and critical phenomena. Oxford Science Publications
38. Bunde A, Havlin S (2012) Fractals and disordered systems. Springer, Berlin
39. Bakke JØH, Hansen A, Kertész J (2006) Failures and avalanches in complex networks. Europhys. Lett. 76(4):717
40. Shang Y (2014) Unveiling robustness and heterogeneity through percolation triggered by random-link breakdown. Phys. Rev. E 90(3):032820
41. Albert R, Jeong H, Barabási A-L (2000) Error and attack tolerance of complex networks. Nature 406(6794):378
42. Cohen R, Erez K, Ben-Avraham D, Havlin S (2000) Resilience of the Internet to random breakdowns. Phys. Rev. Lett. 85(21):4626
43. Callaway DS, Newman ME, Strogatz SH, Watts DJ (2000) Network robustness and fragility: percolation on random graphs. Phys. Rev. Lett. 85(25):5468
44. Sapovadia V (2015) Legal issues in cryptocurrency. In: Handbook of digital currency. Elsevier, Amsterdam, pp 253–266
45. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world'networks. Nature 393(6684):440–442

46. Möser M, Soska K, Heilman E, Lee K, Heffan H, Srivastava S, Hogan K, Hennessey J, Miller A, Narayanan A et al (2018) An empirical analysis of traceability in the monero blockchain. In: Proceedings on privacy enhancing technologies, vol 2018, pp 143–163
47. Krause SM, Danziger MM, Zlatić V (2016) Hidden connectivity in networks with vulnerable classes of nodes. Phys. Rev. X 6(4):041022
48. Krause SM, Danziger MM, Zlatić V (2017) Color-avoiding percolation. Phys. Rev. E 96(2):022313
49. Shekhtman LM, Danziger MM, Bonamassa I, Buldyrev SV, Caldarelli G, Zlatić V, Havlin S (2018) Critical field-exponents for secure message-passing in modular networks. New J. Phys. 20(5):053001
50. Guilbeault D, Becker J, Centola D (2018) Complex contagions: a decade in review. In: Complex spreading phenomena in social systems: influence and contagion in real-world social networks pp 3–25
51. Lin Y, Burghardt K, Rohden M, Noël P-A, D'Souza RM (2018) Self-organization of dragon king failures. Phys. Rev. E 98(2):022127

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.