scientific reports

OPEN

The effect of distant connections on node anonymity in complex networks

Rachel G. de Jong[1,2]✉, Mark P. J. van der Loo[1,2] & Frank W. Takes[1]

Ensuring privacy of individuals is of paramount importance to social network analysis research. Previous work assessed anonymity in a network based on the non-uniqueness of a node's ego network. In this work, we show that this approach does not adequately account for the strong de-anonymizing effect of distant connections. We first propose the use of *d-k-anonymity*, a novel measure that takes knowledge up to distance *d* of a considered node into account. Second, we introduce *anonymity-cascade*, which exploits the so-called infectiousness of uniqueness: mere information about being connected to another unique node can make a given node uniquely identifiable. These two approaches, together with relevant "twin node" processing steps in the underlying graph structure, offer practitioners flexible solutions, tunable in precision and computation time. This enables the assessment of anonymity in large-scale networks with up to millions of nodes and edges. Experiments on graph models and a wide range of real-world networks show drastic decreases in anonymity when connections at distance 2 are considered. Moreover, extending the knowledge beyond the ego network with just one extra link often already decreases overall anonymity by over 50%. These findings have important implications for privacy-aware sharing of sensitive network data.

Network science research[1] is typically about getting a better understanding of the connected structure of a group of people[2], organizations[3], infrastructural objects[4] or other relevant interacting entities[5]. Conducted analyses are often useful for shedding light on societally relevant problems, such as resilience of technical systems[6], predicting systematic financial risk[7], modelling epidemic disease spread[8] or the measurement of socio-economic segregation in a society[9,10]. Crucial for this type of research is the availability of network data representing the interactions that are the object of study. While pseudonymization is often used to mask the identity of individuals in, for example, a social network dataset, the network structure itself may reveal sensitive information on "who is who". As a result, sharing network data imposes risks on the privacy of the entities represented in it. In this paper we set out to discover how we can adequately measure and assess anonymity in complex networks, focusing on methods for discovering how revealing an individual's connections in a network really are.

Related work on privacy in network centers around two major approaches[14–16]: differential privacy[17–19] and k-anonymity[11,20–24]. The first gives randomized answers to user queries such that the privacy of entities are guaranteed, whereas the second enables the sharing of an anonymized version of the network such that there are at least k candidates for each entity. Both of these approaches are strongly embedded in the field of Statistical Disclosure Control (SDC), where traditionally, privacy in relational data is studied[25,26]. However, network data introduces new challenges since the nodes, by which entities are represented, are not isolated observations. Unlike tabular data, the anonymity of a node in a network does not solely depend on the node itself, but can be affected by direct and indirect neighbors in the network. This comes with substantial methodological and computational challenges related to measuring anonymity in network data. A number of different works on anonymity in networks has been published, including various surveys that give a more elaborate overview of this type of work[14–16]. There exist various works on differential privacy, which aim to provide privacy-preserving answers to queries about the network, possibly to eventually generate synthetic network data based on anonymized graph properties[17–19]. Since in this paper we are interested in preventing identity disclosure and ultimately sharing an altered anonymized version of the full network, we have chosen to extend upon the existing line of research[11,20–24] of k-anonymity.

In this paper, we use the notion of k-anonymity and investigate the risk of identity disclosure of nodes based on structural properties of the focal node's surroundings. Noteworthy is that in the remainder of this work, we use the term "anonymity" as a concrete and measurable operationalization of "privacy". We say that a node is

[1]Leiden University, LIACS, 2333 CA Leiden, The Netherlands. [2]Statistics Netherlands, Research and Development, 2492 JP The Hague, The Netherlands. ✉email: r.g.de.jong@liacs.leidenuniv.nl

k-anonymous if there are $k-1$ equivalent nodes in the network according to a particular measure of equivalence. The larger the value of k for a node, the more anonymous the node is. A network as a whole is said to be k-anonymous if all nodes are at least k-anonymous.

As we will see when we turn to our experimental results, in some, but definitely not in all networks does the chosen value of k, ranging from $k = 1$ to 5, strongly affect overall network anonymity beyond $k = 2$. A value of $k = 2$ corresponds to the situation where a node is anonymous if it is not unique based on the employed definition of equivalence. In this particular case, on which we largely focus in this paper, anonymity is effectively equal to non-uniqueness. Various equivalence measures have been used in the literature, taking into account the degree[20], the ego network structure[11,21], the degree distributions of neighboring nodes[22] or the orbits[23,24] of the node under consideration. These measures range from very lenient, accounting for merely the number of connections (degree), to very strict, where nodes are equivalent if they are not distinguishable based on their precise structural position in the network.

All equivalence measures mentioned above correspond to a specific attacker scenario where we assume that someone who tries to de-anonymize entities in the network has a certain type and amount of information. This introduces a trade-off. While using a very strict measure would protect against more attacker scenarios, it would at the same time result in fewer k-anonymous nodes. As a result, when one aims to anonymize the network, e.g., by means of perturbation[21], more changes may be required to ensure that all nodes and therewith the network are k-anonymous. This might have a major impact on the similarity to the original network and hence the so-called utility of the resulting anonymized network. When choosing a measure it is therefore important to account for a realistic amount of attacker information and therewith protect against realistic attacker scenarios. At the same time, the measure should be computable in a reasonable amount of time for nowadays common network sizes of potentially millions of nodes and edges.

In this paper, we aim to contribute to existing literature on this topic in three ways. First, we show the effect on anonymity when one has knowledge beyond the ego network. Empirical results employing the parameterized measure of *d-k-anonymity*[12], for which parameter d denotes the distance from the considered node, indicate that the largest decrease in anonymity occurs when considering 2-neighborhoods (shown in black in Fig. 1B) rather than just the ego networks (1-neighborhoods, shown in red in Fig. 1B). This holds for both well-known graph models and a wide range of real-world networks. Second, we aim to better understand the so-called infectiousness of uniqueness in networks by introducing *anonymity-cascade* (Fig. 1C). This approach extends the aforementioned approach of *d-k-anonymity* by means of a cascading step that finds all nodes that can be uniquely identified if an attacker knows that a particular node is connected to a specific unique node, as illustrated by the pink nodes in Fig. 1C. The newly identified nodes can be reused iteratively, which can result in a cascading effect as illustrated by the orange nodes in Fig. 1C. Our results on a diverse set of real-world networks demonstrate that even knowledge of one extra link, i.e., conducting one cascading "step", frequently reduces overall anonymity by

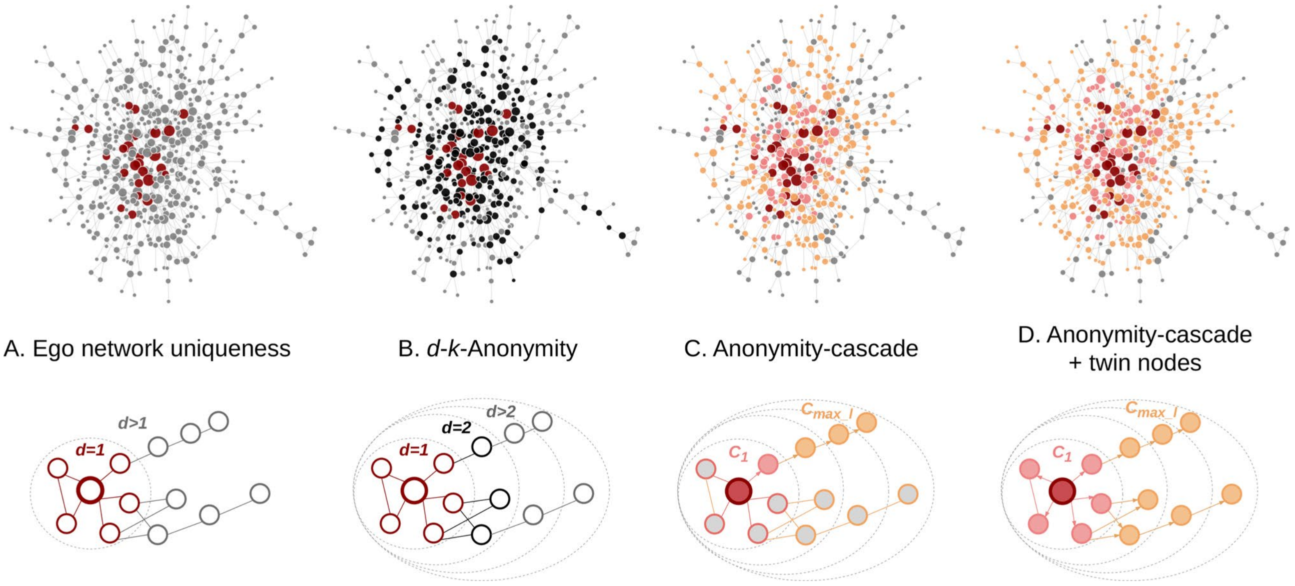

Figure 1. Four approaches for assessing node anonymity: Ego network uniqueness[11] (**A**), followed by the three techniques discussed in this paper: *d-k-anonymity*[12] (**B**), *anonymity-cascade* (**C**) and *anonymity-cascade* with twin nodes (**D**). For each approach, the top row shows the uniquely identified nodes in the giant component of the Copnet calls network[13]. Red nodes are unique using 1-*k-anonymity* (i.e., ego network uniqueness), black nodes with 2-*k-anonymity* (subfigure (**B**) only). Pink nodes can be identified using one cascading step (C_1), orange nodes with multiple steps ($C_{max-\ell}$). Grey nodes are not uniquely identified using the considered approach. The bottom row illustrates an example of a d-neighborhood, detailing which knowledge is taken into account by each approach (edge and node outline color). Subfigure (**C,D**) show the paths traversed by *anonymity-cascade* to identify the pink and orange nodes, given that the red center node is unique for 1-*k-anonymity*.

over 50%. Third, we show how regularities in the underlying graph structure, specifically twin nodes[27], which frequently occur in real-world networks, can be exploited to obtain additional information on certain otherwise indistinguishable entities. This is illustrated in Fig. 1D.

The remainder of this paper is structured as follows. In "Results", we discuss findings resulting from each of the three newly proposed approaches illustrated in Fig. 1B–D, starting with the parameterized anonymity measure and the cascading algorithm. For both, we present results on graph models and real-world networks, before ending with a third and final subsection on the de-anonymizing effect of aforementioned twin nodes. We conclude the paper by summarizing the most important results together with possible directions for future work in "Discussion". Details about the three approaches, the overall experimental setup, code ensuring reproducibility, as well as relevant theorems and proofs, can be found in "Methods".

Results

In this section, we discuss the three approaches to assess node anonymity in a network, each illustrated in Fig. 1. First, in Beyond the ego network, we look at the de-anonymizing effect that knowledge about d-neighborhoods can have when $d \geq 2$. Second, in Anonymity-cascade, we extend *d-k-anonymity* with a cascading step to capture the "infectiousness of uniqueness". For both approaches, we discuss results on both graph models and real-world networks. Third, in Twin nodes, we look at an approach to leverage regularities in the underlying graph structure and identify even more nodes than with the two aforementioned approaches.

Beyond the ego network

In this section, we investigate the effect of knowledge beyond the ego network on anonymity of nodes by using the notion of *d-k-anonymity*[12]. We say that two nodes are d-equivalent if they are indistinguishable with perfect knowledge about their d-neighborhood and position in this neighborhood (see Definition 2 in "Methods"). Here, the d-neighborhood consists of the node itself, all nodes that can be reached by traversing at most d edges, and all edges between these nodes. When $d = 1$, this corresponds to the ego network of the node. This is also illustrated in the bottom of Fig. 1A, B. More precisely, for a pair of d-equivalent nodes the d-neighborhoods are isomorphic (See Definition 1 in "Methods") and their respective position in the d-neighborhood is the same.

If, for a specific node, there are $k - 1$ nodes to which it is d-equivalent, the node is in an equivalence class of size k and we say that the node is d-k-anonymous. If the node is d-1-anonymous, we also call it unique. We summarize the uniqueness of a network as the fraction of unique nodes. Thus, a high network uniqueness implies low anonymity, and low uniqueness implies high anonymity. The results for *d-k-anonymity* are computed by the algorithms described in previous work[12] that builds upon a state-of-the-art isomorphism computation tool[28] (see "Methods" for details). In the following sections, we discuss results of using the measure of *d-k-anonymity* on both graph models and a wide range of real-world networks.

Beyond the ego network in graph models

In this section we investigate the uniqueness of networks (i.e., the fraction of unique nodes) when a possible attacker has perfect knowledge about the ego network or 2-neighborhood of a node. We use three common graph models that each generate networks that reflect a different property that is often observed in real-world networks. For each model, we vary in size and density. The first graph model, the Erdős–Rényi (ER) model[29], generates edges completely at random. Second, the Barabási–Albert (BA) model[30] generates edges by means of the preferential attachment mechanism, which results in the skewed degree distribution that is frequently observed in real-world networks. Third, the Watts–Strogatz (WS) model[31] additionally captures the small world property. More details about, for example, the used parameters, can be found in "Methods".

In Fig. 2 the results on graph models are shown. The figures correspond to the uniqueness maps used by Romanini et al.[11] where the horizontal axis shows the number of nodes, the vertical axis denotes the average degree or m, which equals the number connections made per node for the BA model, or the number of initial connections for each node for the WS model. The color indicates uniqueness of the graph ranging from 0.0 (white, no unique nodes) to 1.0 (dark blue, all nodes unique). Each result is averaged over ten generated graphs.

The results using knowledge of the ego network (top) correspond to the results by Romanini et al.[11] and show a clear connection between the average degree and the fraction of unique nodes. When the number of nodes grows and the average degree is constant, this fraction tends to decrease, meaning that nodes are overall more anonymous. Moreover, these figures show a very clear turning point: for the ranges shown, below the white line almost no nodes are unique while above this line almost all nodes are unique based on their ego network. Results in Supplementary information show that this also holds for higher densities, except when the graph is (near)-complete; for this in real-world networks unrealistic setting, all nodes become non-unique.

However, when we look at the results on the 2-neighborhood computed using *d-k-anonymity* (bottom row of Fig. 2), we see that the uniqueness increases significantly for all models. After an average degree of five almost all nodes are unique, and the uniqueness does not strongly decrease as the network size grows. This shows a large contrast to the results of the 1-neighborhood. Interestingly for $d = 3$ and higher, no large changes occur, which implies that the largest de-anonymizing effect occurs for $d = 2$ (see results up to $d = 5$ in the Supplementary information).

Beyond the ego network in real-world network data

For the next set of experiments, we used a wide range of real-world networks varying in size, density and category. All networks are publicly available and can be found in their corresponding repositories cited in Table 1, which in addition to various experimental results on runtime (further addressed in Anonymity-cascade) summarizes

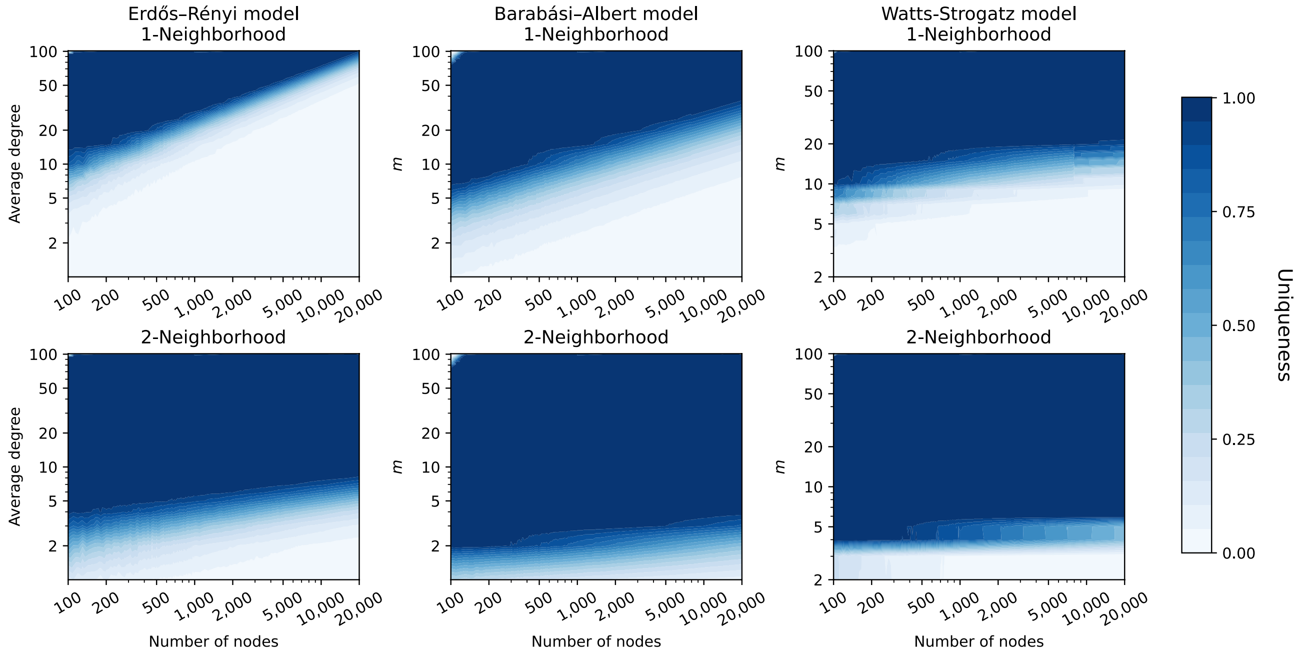

Figure 2. Uniqueness maps using *d*-*k*-anonymity. Maps show network uniqueness, indicated by color, when using information of the 1-neighborhood (top row) and 2-neighborhood (bottom row). Results are shown for the Erdős–Rényi (left), Barabási–Albert (middle) and Watts–Strogatz (right) model with different sizes (horizontal axis) and average degree or *m*, an equivalent thereof (vertical axis).

for each network elementary characteristics such as the number of nodes and edges and the type of network data, covering, e.g., online social networks, co-authorship and biological networks.

In Fig. 3, results are shown for a range of real-world networks for which *d-k-anonymity* could be computed within three hours up to and including $d = 5$[12]. For distance $d = 1$ to $d = 5$ (the five separated columns), we show for $k = 1$ to $k = 5$ by means of color intensity which fraction of the nodes is unique. By reporting on different values for *k*, we aim to take into account that in some cases not being unique does not ensure sufficient privacy, and larger values for *k* are commonly used.

For most networks, similar to previously discussed results on graph models, we observe the largest increase in uniqueness when moving from $d = 1$ to $d = 2$. On average, the absolute increase in uniqueness equals 0.24 with the highest increase being 0.65 for "Moreno health". For 12 out of 26 networks, this additional knowledge more than doubles the number of unique nodes. After $d = 2$ an average increase of 0.04 is observed when increasing to $d = 5$, with the largest value of 0.27 for the "Copnet calls" network. This shows that for most networks, moving from knowledge about the ego network to knowledge about the 2-neighborhood has the largest effect on anonymity. In the figure, we can overall distinguish between three different cases: (1) a high uniqueness at $d = 1$, or if there is a low uniqueness at $d = 1$, there is either (2) a high uniqueness at $d = 2$, or (3) a low uniqueness at all distances. In Supplementary information, we include results aiming to correlate the uniqueness found to various graph properties. For the networks presented in Table 1, we find that networks with a larger diameter or average path length tend to have a lower uniqueness. Networks with high degrees or density tend to have a higher uniqueness. This is also shown by the work of Romanini et al.[11] and is similar to earlier results obtained for the graph models.

Additionally, we compare different values for *k*, where we measure the fraction of nodes that are at most *k*-anonymous, and hence the fraction of nodes for which there are at most *k* candidates for an attacker with knowledge of their *d*-neighborhood. Increasing the value for *k* to 5 results in an average increase of 0.05 to 0.08 with the largest increase equal to 0.33 for the "Moreno innovation" communication network. The results show that in many cases larger values beyond $k = 2$ up to $k = 5$ do not result in a large decrease in anonymity. Thus, we can learn a lot by only distinguishing between unique and non-unique nodes to measure anonymity. With this in mind, and in the interest of readability of further results, we choose to report on uniqueness ($k = 1$) and *d-k-anonymity* with $d = 1$ and $d = 2$ in the remainder of this paper. Overall, we conclude that accounting for knowledge beyond the ego network in both graph models and real-world networks shows a significant decrease in node anonymity. For completeness, a specific figure showing the uniqueness using different values for *d* can be found in Supplementary information.

Anonymity-cascade

The measure of *d-k-anonymity* employed above, while more informative than ego network uniqeueness, has two noteworthy disadvantages. First, due to isomorphism computations of possibly large neighborhoods, for larger values of *d*, this approach is computationally expensive[12] (see also the runtimes for $d = 2$ in the seventh column of Table 1). Second, knowledge of the 2-neighborhoods can be an unrealistic attacker scenario; in particular if

Network	Nodes	Edges	Fraction twins	$max - \ell$	Runtime *d-k-anonymity*		Runtime *Anonymity-cascade*		Type
					$d = 1$	$d = 2$	C_1	$C_{max-\ell}$	
Radoslaw emails[32]	167	3250	0.072	3	0.02 s	+ < 0.01 s	+ < 0.01 s	+ < 0.01 s	Communication
Primary school[33]	236	5899	0.000	1	0.02 s	+ < 0.01 s	+ < 0.01 s	+ < 0.01 s	Human contact
Moreno innov.[32]	241	923	0.025	3	< 0.01 s	+ < 0.01 s	+ < 0.01 s	+ < 0.01 s	Communication
Gene fusion[32]	291	279	0.753	6	< 0.01 s	+ < 0.01 s	+ < 0.01 s	+ < 0.01 s	Biological
Copnet calls[13]	536	621	0.287	10	< 0.01 s	+ 0.01 s	+ < 0.01 s	+ < 0.01 s	Communication
Copnet sms[13]	568	697	0.285	7	< 0.01 s	+ 0.01 s	+ < 0.01 s	+ < 0.01 s	Communication
Copnet FB[13]	800	6418	0.005	4	0.02 s	+ 0.01 s	+ < 0.01 s	+ < 0.01 s	Online social
FB Reed98[34]	962	18,812	0.012	3	0.05 s	+ 0.01 s	+ 0.01 s	+ 0.02 s	Online social
Arenas email[32]	1133	5451	0.042	5	0.02 s	+ 0.01 s	+ < 0.01 s	+ < 0.01 s	Communication
Network science[32]	1461	2742	0.755	6	0.01 s	+ 0.01 s	+ < 0.01 s	+ < 0.01 s	Co-autorship
FB Simmons81[34]	1518	32,988	0.011	3	0.12 s	+ 0.02 s	+ 0.01 s	+ 0.02 s	Online social
DNC emails[32]	1893	4385	0.706	4	0.02 s	+ 0.38 s	+ < 0.01 s	+ < 0.01 s	Online social
Moreno health[32]	2539	10,455	0.003	5	0.03 s	+ 0.04 s	+ < 0.01 s	+ < 0.01 s	Human social
FB Wellesley22[34]	2970	94,899	0.000	4	0.4 s	+ 0.07 s	+ 0.04 s	+ 0.08 s	Online social
Bitcoin alpha[34]	3783	14,124	0.306	5	0.04 s	+ 0.36 s	+ < 0.01 s	+ 0.01 s	Online social (trust)
GRQC collab.[35]	5242	14,484	0.455	8	0.03 s	+ 0.04 s	+ < 0.01 s	+ < 0.01 s	Co-autorship
FB Carnegie49[34]	6637	249,967	0.008	4	1.37 s	+ 0.44 s	+ 0.09 s	+ 0.18 s	Online social
Pajek Erdős[32]	6927	11,850	0.737	6	0.02 s	+ 0.19 s	+ 0.01 s	+ 0.02 s	Co-autorship
DT interaction[36]	7341	15,138	0.572	12	0.07 s	+ 2.88 s	+ < 0.01 s	+ < 0.01 s	Biological
DG assoc.[36]	7813	21,357	0.531	8	0.21 s	+ 5.13 s	+ < 0.01 s	+ < 0.01 s	Biological
FB GWU54[34]	12,193	469,528	0.006	4	2.64 s	+ 0.92 s	+ 0.18 s	+ 0.36 s	Online social
Anybeat[34]	12,645	49,132	0.500	5	0.41 s	+ 1.34 h	+ 0.02 s	+ 0.04 s	Online social
CE-CX[34]	15,229	245,952	0.021	6	1.04 s	+ 1.56 s	+ 0.09 s	+ 0.19 s	Biological
Astro Physics[34]	18,771	198,050	0.305	6	0.72 s	+ 1.91 s	+ 0.05 s	+ 0.11 s	Co-autorship
FB BU10[34]	19,700	637,528	0.006	4	3.17 s	+ 1.44 s	+ 0.25 s	+ 0.50 s	Online social
FB Uillinois[34]	30,664	1,048,574	0.002	4	5.96 s	+ 2.66 s	+ 0.39 s	+ 0.78 s	Online social
Enron email[35]	36,692	183,831	0.528	6	0.90 s	+ 1.44 m	+ 0.06 s	+ 0.12 s	Communication
FB Penn94[34]	41,536	1,362,220	0.002	4	8.95 s	+ 5.39 s	+ 0.50 s	+ 1.00 s	Online social
FB wall 2009[32]	46,952	183,412	0.133	8	0.54 s	+ 1.80 s	+ 0.05 s	+ 0.13 s	Communication
Brightkite[34]	58,228	214,078	0.258	8	0.78 s	+ 18.41 s	+ 0.06 s	+ 0.14 s	Online social
The marker cafe[37]	69,413	1,644,843	0.200	5	37.96 s	+ 10.46 m	+ 0.67 s	+ 1.35 s	Human contact
Slashdot zoo[32]	79,116	467,731	0.274	7	3.36 s	+ 2.62 m	+ 0.15 s	+ 0.34 s	Online social
Twitter[32]	465,017	833,540	0.801	5	59.53 s	+ 6.04 h	+ 0.29 s	+ 0.62 s	Online social
DBLP[32]	1,824,701	8,344,615	0.402	10	51.37 s	+ 11.58 m	+ 28.1 s	+ 57.18 s	Co-autorship
Flixster[32]	2,523,386	7,918,801	0.631	8	3.10 m	+ 8.57 h	+ 18.58 s	+ 37.81 s	Online social
Youtube[32]	3,223,585	9,375,374	0.336	15	13.24 m	+ > 1 week	+ 30.68 s	+ 1.04 m	Online social

Table 1. Overview of the real-world networks used in the experiments. For each network, we list the number of nodes, edges, fraction of twin nodes, the highest attained cascading level ($max - \ell$) and runtimes of the experiments performed containing the total runtime ($d = 1$) and runtime additional to computing $d = 1$, indicated by "+" for $d = 2$, C_1 and $C_{max-\ell}$.

the 2-neighborhood has a complex structure. However, it is not unreasonable to assume that an attacker obtains some information beyond the ego network, especially if the 1-neighborhood is small or the 2-neighborhood sparse. If that knowledge includes that the node is connected to a unique node, which can be concluded using knowledge of the 1-neighborhood, this may strongly decrease the number of candidates. In some cases this can be sufficient to uniquely identify a node.

We propose to explicitly detect this by introducing *anonymity-cascade* (C_ℓ), an algorithm that extends *d-k-anonymity* and accounts for the so-called "infectiousness of uniqueness". We assume that an attacker has knowledge about the 1-neighborhoods of two nodes, of which one is unique, and that there is a connection between them. The *anonymity-cascade* algorithm starts by finding all nodes that can be uniquely identified by knowing that this node 1) is connected to a unique node u using 1-*k-anonymity* and 2) is unique in the set of neighbors of node u.

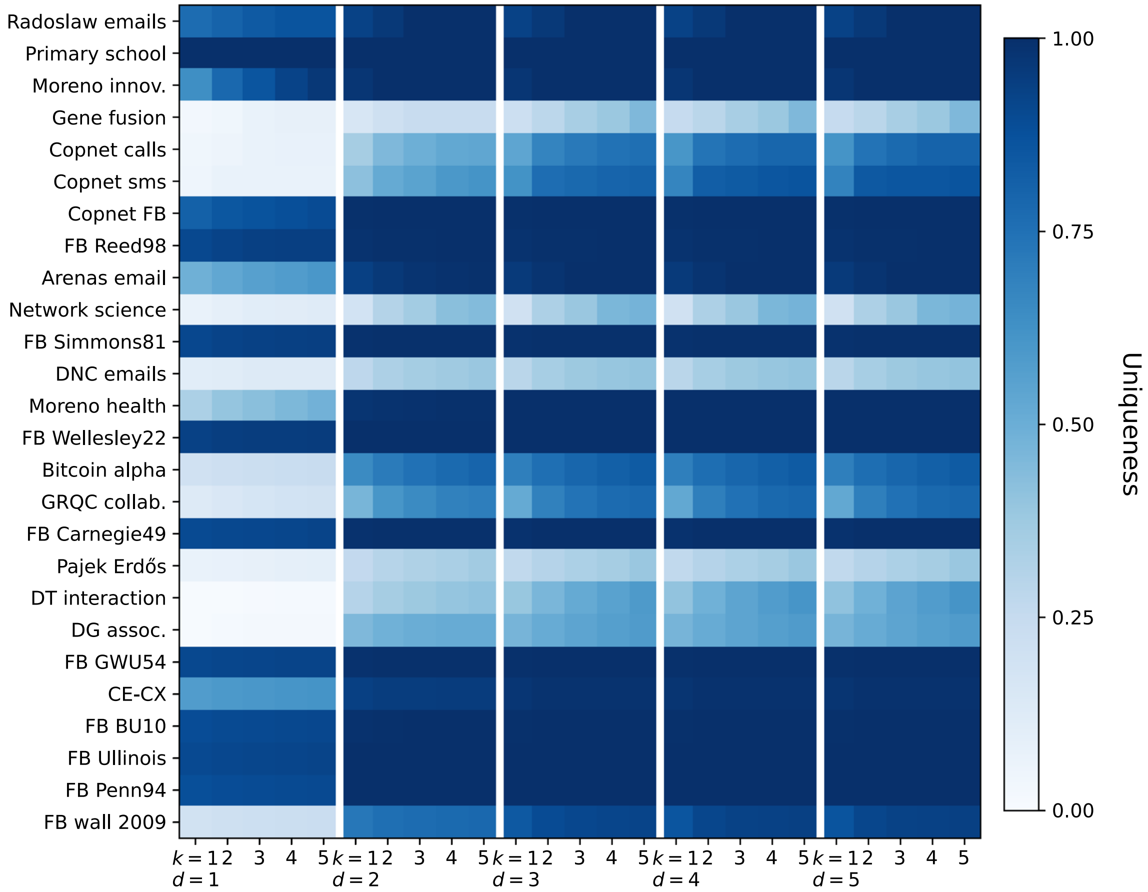

Figure 3. *d*-*k*-Anonymity in real-world networks. Results are shown for for the 29 real-world networks in Table 1 for which *d-k-anonymity* with $d = 5$ could be computed within 3 h. Each cell denotes the fraction [ranging from 0.0 (white) to 1.0 (dark blue)] of nodes that are $\leq k$-anonymous when accounting for knowledge of the *d*-neighborhood.

We refer to this as the first level of the cascading algorithm, denoted C_1. The nodes identified with C_1 can then be used to continue this cascading effect among further levels in a Breadth-First Search manner for ℓ steps or "levels", as illustrated in Fig. 1C. In this figure, the process starts at the unique red node. From this node, the pink node can be uniquely identified (C_1) and from there the nodes can be used to identify the orange nodes ($C_{\geq 2}$). This process can be repeated until no more unique nodes are found, the then attained level is called $max - \ell$. We refer to this as cascading final, denoted $C_{max-\ell}$. While going beyond the first level mimics a less realistic attacker scenario, this approach gives insights into how far the cascading effect can continue (shown in the fifth column of Table 1), and the effect this can potentially have.

For one level of cascading (C_1), the measure is more strict than 1-*k-anonymity* and at most as strict as 2-*k-anonymity*, which is explained in Theorem 1 and Proof 1 in "Methods". Additionally, *anonymity-cascade* is less expensive to compute than 2-*k-anonymity* and allows us to assess anonymity in larger networks with millions of nodes, as shown in Table 1. A more detailed description of *anonymity-cascade* can be found in "Methods".

Anonymity-cascade in graph models

In Fig. 4, the results of *anonymity-cascade* on graph models can be found, similar to the uniqueness maps in Fig. 2. The top row shows the results of C_1 (one level of cascading), which resemble the results of *d-k-anonymity* with $d = 1$. Apparently, for these graph models, having knowledge about one additional link does not strongly affect the overall anonymity. When the algorithm continues up to its final level ($C_{max-\ell}$), as shown in the bottom row, the results change significantly and many more nodes are unique.

The maximal depth reached by *anonymity-cascade* ($max - \ell$) can be very high; averages of 38, 14 and 15 are observed for ER, BA and WS respectively. Especially for sparse graphs with over 15,000 nodes high values are observed (results can be found in Supplementary information). Figure 4 also shows that for $C_{max-\ell}$ the fraction of unique nodes is more stable when the graph size increases compared to 2-*k-anonymity* in Fig. 2 for the ER and WS models. This seems to indicate that in graph models, which are more random than real-world networks, cascading has a small but local effect that can continue for many levels, especially in large graphs.

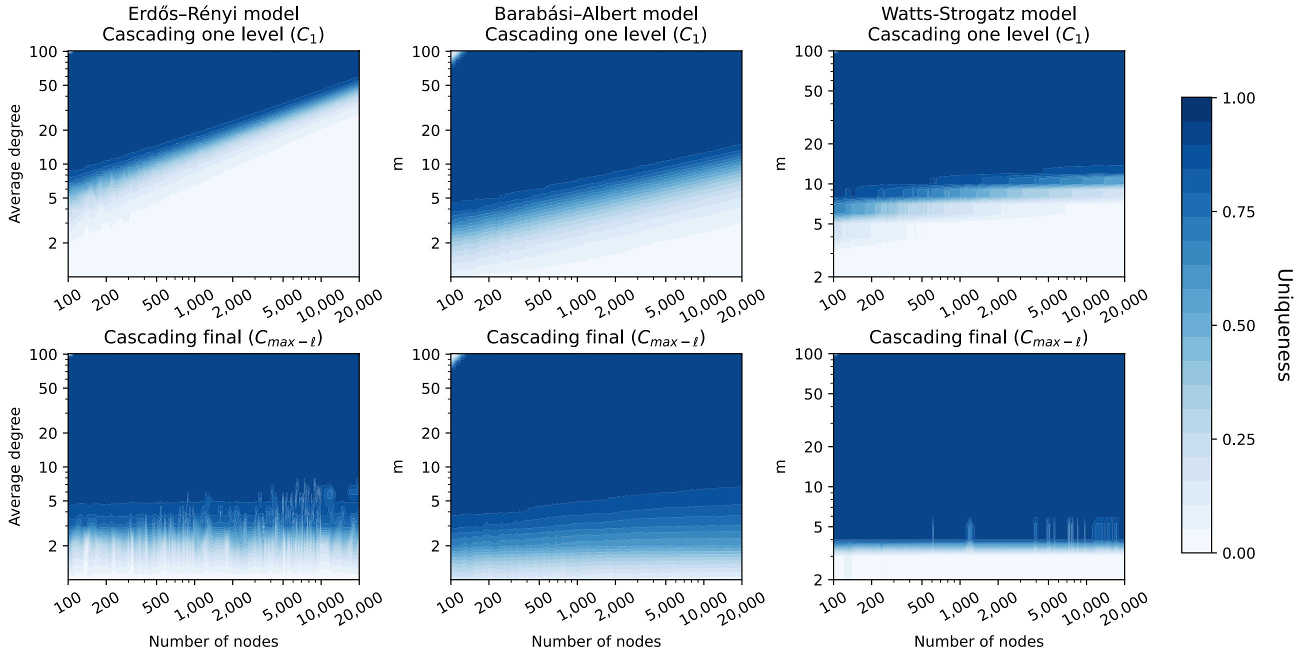

Figure 4. Uniqueness maps using *anonymity-cascade*. Maps show network uniqueness, indicated by color, when using one level of cascading (top row) and up to the final level of cascading (bottom row). Results are shown for the Erdős–Rényi (left), Barabási–Albert (middle) and Watts–Strogatz (right) model with different sizes (horizontal axis) and average degree or *m*, an equivalent thereof (vertical axis).

Anonymity-cascade in real-world network data

The results in Fig. 5 show the fraction of unique nodes when using *anonymity-cascade* (C_1 and $C_{max-\ell}$) compared to *d-k-anonymity* with $d = 1$ and $d = 2$. The results show that when having additional knowledge of one extra link, the fraction of uniquely identifiable nodes doubles for 19 out of 36 networks. The largest *anonymity-cascade* increase equals 0.5 for "Moreno health": while at $d = 1$ a fraction of 0.33 is uniquely identified, one level of cascading results in a uniqueness of 0.83.

The results additionally show that the fractions obtained by C_1 are often close to the fraction identified by 2-*k-anonymity*. On average the difference equals just 0.09. However, for some networks this difference is larger: for 13 networks this fraction is still larger than 0.1. Overall, while the runtimes in Table 1 showed that computing 2-*k-anonymity* can be computationally expensive, especially in large networks, using *anonymity-cascade* provides an adequate estimate of 2-*k-anonymity* in a reasonable amount of time, even for networks with millions of nodes and edges.

When continuing the cascading effect until the the final level ($C_{max-\ell}$), this results in an additional increase in uniqueness for many networks. While this level of knowledge is less realistic as an attacker scenario, this gives insights into how far one could exploit this cascading effect. For five of the networks, this results in a uniqueness increase larger than a factor two. The cascading approach can hence be effective, even at higher levels. As shown in Table 1, the value for $max - \ell$ achieved can differ per network. In Supplementary information, we include results aiming to explain the difference. Large average path lengths and diameter seem to result in longer possible cascading paths for these networks, which intuitively makes sense given that the value is based on paths. Higher degrees and densities are overall likely to result in shorter paths. However, contrary to what the results for graph

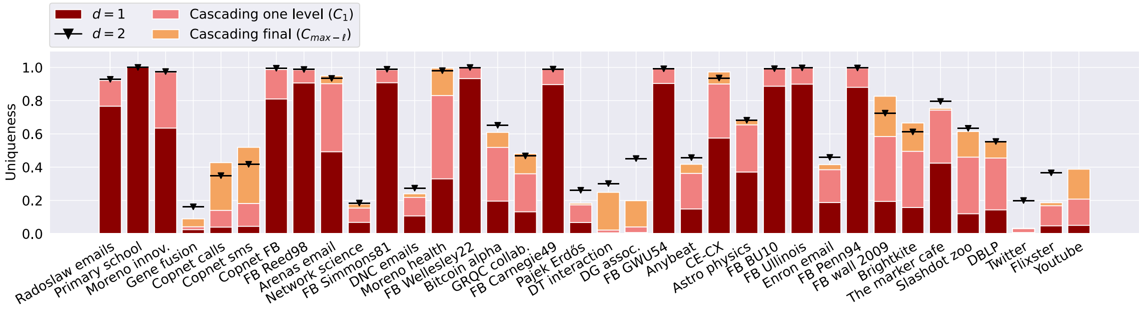

Figure 5. Uniqueness in real-world networks. Fraction of unique nodes (vertical axis) on different datasets (horizontal axis) when accounting for different levels of information: 1-neighborhood (red), 2-neighborhood (black line with triangle), cascading one level (pink) and the cascading final (yellow).

models show, for most real-world networks the largest increase in uniqueness happens at C_1 (see Supplementary information for a figure detailing the drastically decreasing effect per subsequent level).

Twin nodes

The previous two approaches can be extended by means of a "twin node"[27] processing step. This concerns the identification of sets of nodes that all share the exact same neighbors which we refer to as twin nodes. Focusing on the example of two pairs of such nodes in Fig. 6B-C, we distinguish two cases: either the nodes are connected to the same nodes (open twin nodes, Fig. 6B), or they additionally have a connection between each other (closed twin nodes, Fig. 6C). This differs from the case illustrated in Fig. 6A. If a full network is shared in a pseudonymized format, such as the networks listed in Table 1, then sets of twin nodes can be derived from the network itself, without requiring any additional external information. It turns out that twin nodes can occur frequently in real-world networks: in Table 1, fractions up to 0.801 are observed.

In Proof 2, we show that twin nodes are indistinguishable based on any structural property, implying that any structure-based measure for equivalence can not distinguish between these nodes. As a result, if a node has at least one twin, it can not be unique. At the same time, if in an attempt to identify entities all candidates for a node are twin, then this gives the same information as when the node is uniquely identified in the network.

The reason is that for these nodes we know both their exact structural position in the network and which nodes they are connected to, which is the same information if a node is uniquely identified. This notion relates to group disclosure in SDC[38]: even if there are multiple candidates for an entity, it might still be possible to derive sensitive information if all candidates have the same attribute. In our situation, their structural position in the network and the connections of the node. In SDC, the problem is overcome by introducing the notion of ℓ-diversity[39] which extends k-anonymity with the requirement that for each of the k candidates, there should be at least ℓ different sensitive values.

We process twin nodes in our approaches as follows. First, for *d-k-anonymity*, we say that a node is *twin-unique* if either the node is unique, or all candidates for the node are twins of each other. Second, in *anonymity-cascade*, we start with all nodes that are twin-unique using 1-*k-anonymity*. If in the cascading step all candidates are twins, they are also twin-unique, and we continue the cascading process from each of the nodes (see also the example in Fig. 1D).

Twin nodes in real-world network data

Figure 7 shows the fraction of twin-unique nodes for the considered corpus of real-world networks, extending upon Fig. 5. Accounting for twin-unique nodes in one level of *anonymity-cascade* results in an average absolute increase of 0.16, compared to uniqueness. For the "Twitter" network the largest difference is observed resulting in a twin-uniqueness of 0.65. An explanation for this strong effect is in the high fraction of twin nodes in the network (0.801, see Table 1). These types of nodes are likely frequently occurring due to the preferential attachment process often observed in these types of social networks[1]. This results in many nodes with a low degree, which would be identified using one cascading step if their neighbor is unique. Overall, the relative increase is larger if the network has a higher fraction of twin nodes (we detail this relation in Supplementary information).

For cascading final and *d-k-anonymity* with $d = 2$, we observe similar average increases of 0.21 and 0.18 respectively. The largest increase is again found for the "Twitter" network and equals 0.76 and 0.69. In contrast, merely accounting for twin nodes using the plain measure of 1-*k-anonymity* almost never increases the fraction; differences of at most 0.04 are observed. This small effect is likely due to the fact that there are often multiple sets of twin nodes with the same ego network implying that these nodes often have more frequently occurring, likely simple, ego networks.

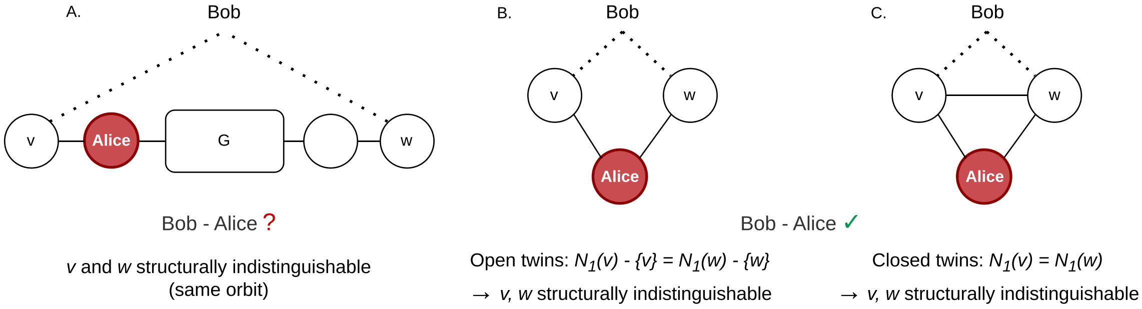
A.
Bob
v
Alice
G
w
Bob - Alice ?
v and w structurally indistinguishable
(same orbit)
B.
Bob
v
w
Alice
Bob - Alice ✓
Open twins: $N_1(v)$ - {v} = $N_1(w)$ - {w}
→ v, w structurally indistinguishable
C.
Bob
v
w
Alice
Closed twins: $N_1(v) = N_1(w)$
→ v, w structurally indistinguishable

Figure 6. Illustration of two types of structurally indistinguishable nodes. (**A**) Two candidate nodes for Bob (indicated by dotted lines) are structurally indistinguishable, but do not share their connections. Given that the position of Alice in the network is known (e.g. using 1-*k-anonymity*, there is no certainty about the connection between Bob and Alice. (**B,C**) All candidates for Bob are twin-unique. The nodes are structurally indistinguishable (see Theorem 2 and Proof 2) and an attacker can be certain about the connections of Bob and thus the connection with Alice.

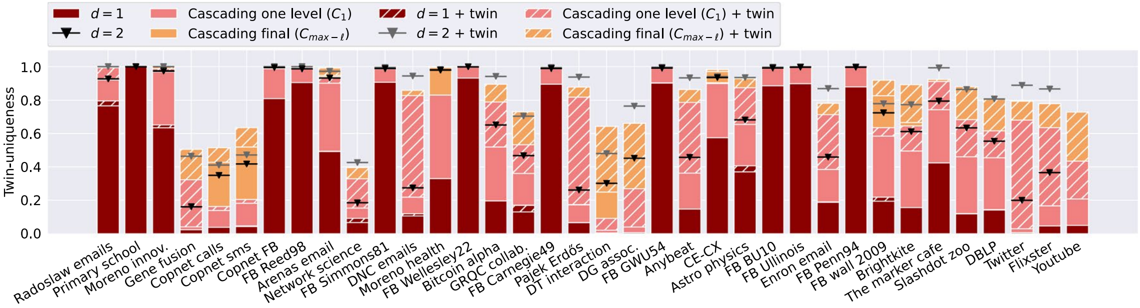

Figure 7. Twin-uniqueness in real-world networks. Fraction of (twin)-unique nodes (vertical axis) on different datasets (horizontal axis) when accounting for different levels of information: 1-neighborhood (red), 2-neighborhood (black and grey line with triangle), one level of cascading (pink) and all levels of cascading (yellow). Results without hatching correspond to results in Fig. 5. Results with hatching show the increase achieved by including twin-unique nodes.

Discussion

In this work we discussed new measures and algorithms for anonymity which each correspond to a different attacker scenario, i.e., amount of knowledge of the network structure, and vary in strictness and required computational effort.

First, we explored *d-k-anonymity* and found that compared to merely measuring ego network uniqueness, information about 2-neighborhoods drastically decreased the anonymity of nodes in both graph models and a wide range of real-world networks. However, perfect knowledge of the 2-neighborhood might in some cases be an unrealistically large amount of knowledge, especially if the ego network is dense. Therefore, we extended the measure of *d-k-anonymity* by means of a cascading step, resulting in *anonymity-cascade*, which models the scenario in which a possible attacker has knowledge about ego networks of two linked nodes, of which one is unique. *Anonymity-cascade* adds to existing measures in two ways. First, it accounts for the scenario where a possible attacker has more information than the ego network, but less than the 2-neighborhood. Second, since *anonymity-cascade* is far less expensive to compute, it can effectively measure anonymity on larger networks with millions of nodes and over ten million edges in minutes.

When assuming above mentioned additional knowledge of one link, we observed major increases in the uniqueness (and thus decreases in anonymity) of virtually all of the considered real-world network datasets. For 19 of the investigated networks, using one level of cascading more than doubled the fraction of unique nodes. When continuing this cascading effect, the anonymity decreased further for some networks, but the effect of the first link remains the largest. Based on results using *d-k-anonymity* and *anonymity-cascade*, we argue the de-anonymizing effects can not be ignored, and that, as opposed to what was assumed in previous work, information beyond the ego network should be taken into account when measuring anonymity of individuals in networks.

Lastly, we discussed that if all candidates for a node are connected to the exact same nodes, i.e., are twin nodes, the same knowledge can be obtained as when a node is uniquely identified, without any additional attacker knowledge. Taking into account the notion of twin-uniqueness, which includes all nodes for which all candidates are twin, we observed a large decrease in anonymity for many of the real-world networks.

With this work, we aim to emphasize the importance of research on anonymity measures for networks. But moreover, we wish to stress the relevance of balancing the trade-off between attacker knowledge, measure strictness and computation time. While we have now arrived at an approach that can process large networks with millions of nodes and edges in a reasonable amount of time, there are still many directions for possible future work. One avenue for future work could be to explore variants of the cascading measure, using a different starting knowledge than the uniqueness of the ego network. Another avenue of research could be to investigate how other graph properties can be exploited, such as distance between candidate nodes, or their membership of network communities. Another way to extend the work would be to include other properties such as node or edge labels, weights or timestamps. In the end, the aim is to create feasible measures for anonymity that take into account realistic attacker scenarios, while remaining computable in a realistic amount of time. Moreover, based on all of these measures, network anonymization methods can be created and applied, ultimately allowing researchers and practitioners to share network data with the assurance that the privacy of individuals in it is maximally guaranteed.

Methods

In this section, we discuss relevant definitions, theorems and proofs as well as a detailed description of the proposed algorithms, ending with a description of the overall experimental setup.

Notation and definitions

We define a network or graph $G = (V, E)$ as a set of nodes V and set of edges $\{u, v\} \in E$, where $u, v \in V$. The set of all nodes in a particular graph G is denoted $V(G)$. Given two nodes $v, w \in V$ we define the distance $d(v, w)$ as the minimum number of adjacent edges that must be traversed to from node v reach w. If there is no such path, $d(v, w) = \infty$. It follows that $d(v, v) = 0$, and since the graph is undirected, $d(v, w) = d(w, v)$. For a node v, we define the d-neighborhood $N_d(v)$ as the graph containing all nodes that are at most distance d from v, and the set of all edges between these nodes.

The *d*-*k*-anonymity algorithm

To measure anonymity, we use d-k-anonymity[12], an existing approach that uses isomorphism. We follow notation and definitions used in[12]. This measure takes as input a graph and a value for the neighborhood distance d, and outputs an equivalence partition of the nodes such that in each equivalence class, all nodes are equivalent to each other. Based on this partition, the fraction of nodes that are k-anonymous can be determined for any given value k. Recall that a node is k anonymous if it is equivalent to $k-1$ nodes. If $k=1$ for a node, it is unique. If a node is in an equivalence class of size k, i.e., it is equivalent to $k-1$ nodes, we say that it is k-anonymous.

Definition 1 Graph isomorphism. Given two graphs $G=(V,E)$ and $G'=(V',E')$, a graph isomorphism is a bijective function $\phi : V \to V'$ such that for each $v, w \in V$ it holds that $\{\phi(v), \phi(w)\} \in E'$ precisely when $\{v, w\} \in E$.

We call two graphs isomorphic if there is at least one isomorphism between them. A special form of isomorphism is an automorphism, denoted by γ. This is an isomorphism from a graph onto itself. If there exists an automorphism such that two nodes $v, w \in V$ are mapped onto each other, they are said to be in the same orbit.

Definition 2 *d-Equivalence*. Two nodes $v, w \in V$ are d-equivalent if: (1) their d-neighborhoods are isomorphic and (2) there is an isomorphism ϕ such that $\phi(v)=w$.

If two nodes are d-equivalent, they are equivalent with respect to *d*-*k*-*anonymity*. It is not difficult to show that d-equivalence indeed satisfies the properties of an equivalence relation[40] (identity, reflexivity and transitivity). It can also be demonstrated that if two nodes are $d+1$-equivalent, they must also be d-equivalent. In previous work[12] several methods for efficient calculation of this measure, such as the use of a cache and filtering out nodes based on graph invariants, are discussed. Code for computing d-k-anonymity can be found at https://github.com/RacheldeJong/dkAnonymity.

The anonymity-cascade algorithm

Below we describe the workings of *anonymity-cascade*, the algorithm proposed in this paper that extends *d*-*k*-*anonymity* with $\ell \geq 1$ cascading steps. It takes as input a graph and depth parameter ℓ, and outputs the uniquely identified nodes.

```
• Input: Graph $G=(V,E)$, maximum cascading level $\ell$
• $U_{cur} = U =$ all nodes with a unique ego network (obtained using d-k-anonymity with $k=1$ and $d=1$)
• $\ell_{cur} = 1$
• While $\ell_{cur} \leq \ell$ and $U_{cur} \neq \emptyset$:
    – $U_{new} = \emptyset$
    – For each node $u \in U_{cur}$:
        * For each $v \in V(N_1(u)) - \{u\}$:
            · Check if there is a node $v' \in V(N_1(u)) - \{u, v\}$ such that $v$ and $v'$ are 1-equivalent
            · If there is no such node: $U_{new} = U_{new} \cup \{v\}$
    – $U_{cur} = U_{new} - U$
    – $U = U \cup U_{new}$
    – $\ell_{cur} = \ell_{cur} + 1$
• Output: Set of uniquely identified nodes $U$
```

Interestingly, *anonymity-cascade* has the useful property that all nodes that are unique for C_1, are also unique for 2-*k*-*anonymity* and hence 2-*k*-*anonymity* is at least as strict as C_1 (see Theorem 1 and Proof 1). We use a proof by contradiction and first assume that there is a pair of nodes v, w that is 2-equivalent and v is unique for C_1. We then show a unique neighborhood is contained in the 2-neighborhood of both v and w, which contradicts the assumption that v is unique for C_1. Code for computing *anonymity-cascade* can be found at https://github.com/RacheldeJong/Anonymitycascade.

Theorem 1 *In a graph $G=(V,E)$, a node v that is unique using anonymity-cascade with $\ell=1$ (C_1) is also unique using 2-k-anonymity.*

Proof Node v is connected to a node u where $N_1(u)$ is unique. It holds that $N_1(u)$ is contained in $N_2(v)$, as $d(v,u') \leq 2$ for all $u' \in V(N_1(u))$. Let us now assume there is a node w which is 2-equivalent to v. This implies that $N_2(v)$ and $N_2(w)$ are isomorphic. Hence, $N_1(u)$ should occur in $N_2(w)$. Since $N_1(u)$ is unique, this implies w

should be a neighbor of u. However, v and w can only be 2-equivalent if they are 1-equivalent. This contradicts the assumption that v is unique using C_1. □

Twin node preprocessing

For the definition of twin nodes, we distinguish between a closed neighborhood $N_d(v)$ as the neighborhood defined in Notation and definitions and the open neighborhood of a node $N'_d(v) = N_d(v) - \{v\}$.

Definition 3 Twin nodes. Given a graph $G = (V, E)$, nodes $v_1 \neq v_2 \in V$ are closed twin nodes if $N_1(v_1) = N_1(v_2)$ or open twin nodes if $N'_1(v_1) = N'_1(v_2)$.

Twin nodes are in the same orbit, which we show by constructing an automorphism that maps the twin nodes onto each other, and all other nodes onto themselves. The proof used is similar to the proof presented in previous work[12].

Theorem 2 *Given a graph $G = (V, E)$ and nodes $v_1, v_2 \in V$. If nodes v_1, v_2 are twin nodes, then they are in the same orbit.*

Proof Given twin nodes v_1 and v_2, we define an automorphism $\gamma : V \rightarrow V$ that swaps v_1 and v_2 and maps all other nodes onto themselves. To show that γ is a valid automorphism, note that for any $\{v_1, w \neq v_2\} \in E$, we have $\{\gamma(v_1), \gamma(w)\} = \{v_2, w\} \in E$ because of the twin node property (Definition 3). Similarly we have $\{\gamma(v_2), \gamma(w)\} = \{v_1, w\} \in E$ for all $\{v_2, w \neq v_1\} \in E$. If $\{v_1, v_2\} \in E$, then $\{\gamma(v_1), \gamma(v_2)\} = \{v_1, v_2\} \in E$. Hence, v_1 and v_2 are in the same orbit. □

We find twin nodes using the algorithm below, which serves as a preprocessing step before *d-k-anonymity* or *anonymity-cascade* is executed. It is worth to note that a node can be a twin of more than one node, and hence a set of twin nodes can have a size larger than two.

- **Input:** Graph $G = (V, E)$
- Create two dictionaries M_o and M_c, used to map neighborhoods onto nodes with this resp. open and closed neighborhood
- For each node $v \in V$:
    - If $V(N'_1(v)) \in M_o$: $M_o[V(N'_1(v))] = M_o[V(N'_1(v))] \cup \{v\}$
    - Else if $V(N_1(v)) \in M_c$: $M_c[V(N_1(v))] = M_c[V(N_1(v))] \cup \{v\}$
    - Else: $M_o[V(N'_1(v))] = \{v\}$ and $M_c[V(N_1(v))] = \{v\}$
- **Output:** Sets of open and closed twin nodes M_o, M_c

After finding twin nodes, they are taken into account as follows. When computing *d-k-anonymity*, we select one node for each set of twin nodes that has to be taken into account when computing anonymity. All other twin nodes are not taken into account during computation, and afterwards added to their corresponding equivalence classes[12]. When computing twin-uniqueness for both *d-k-anonymity* and *anonymity-cascade*, both unique nodes and nodes for which all candidates are twins are twin-unique. For *anonymity-cascade*, we additionally reuse twin-unique nodes to continue the cascading effect.

Experimental setup

The results for *d-k-anonymity* and *anonymity-cascade* were obtained using the default settings of the code repositories linked above. When computing twin-uniqueness, for both *d-k-anonymity* and *anonymity-cascade* we compute and take into account both open and closed twin nodes; these computation times are not included in reported runtimes for *d-k-anonymity*, and are included for runtimes for *anonymity-cascade*.

The graph models used for the experiments consist of the Erdős–Rényi[29], Barabási–Albert[30], and Watts–Strogatz[31] models generated using NetworkX[41]. For the WS graphs we use random wiring probability $p_r = 0.5$, similar to previous work[11]. All results reported on graph models are averaged over ten networks.

For experiments on real-world networks, self loops and nodes without edges are removed from the original network dataset. Moreover, edge weights, timestamps and directionality are ignored. Reported runtimes are averaged over five runs. All experiments are conducted on a machine with 1 TB RAM, 64 AMD EPYC 7601 cores, and 128 threads. During the experiments, each run uses one thread, which is not shared with other processes.

Data availibility

All network datasets are available in the repositories cited in Table 1.

Received: 23 June 2023; Accepted: 22 December 2023
Published online: 12 January 2024

References

1. Barabási, A. L. *Network Science* (Cambridge University Press, 2016).
2. Bokányi, E., Heemskerk, E. M. & Takes, F. W. The anatomy of a population-scale social network. *Sci. Rep.* **13**, 9209 (2023).

3. Garcia-Bernardo, J., Fichtner, J., Takes, F. W. & Heemskerk, E. M. Uncovering offshore financial centers: Conduits and sinks in the global corporate ownership network. *Sci. Rep.* **7**, 1–10 (2017).
4. Guimera, R. & Amaral, L. A. N. Modeling the world-wide airport network. *Eur. Phys. J. B* **38**, 381–385 (2004).
5. Barabasi, A.-L. & Oltvai, Z. N. Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
6. Ganin, A. A. *et al.* Operational resilience: Concepts, design and analysis. *Sci. Rep.* **6**, 1–12 (2016).
7. Cimini, G., Squartini, T., Garlaschelli, D. & Gabrielli, A. Systemic risk analysis on reconstructed economic and financial networks. *Sci. Rep.* **5**, 1–12 (2015).
8. Azizi, A., Montalvo, C., Espinoza, B., Kang, Y. & Castillo-Chavez, C. Epidemics on networks: Reducing disease transmission using health emergency declarations and peer communication. *Infect. Dis. Model.* **5**, 12–22 (2020).
9. Bojanowski, M. & Corten, R. Measuring segregation in social networks. *Soc. Netw.* **39**, 14–32 (2014).
10. Kazmina, Y., Heemskerk, E. M., Bokanyi, E. & Takes, F. W. Socio-economic segregation in a population-scale social network. *arXiv preprint*arXiv:2305.02062 *(2023).*
11. Romanini, D., Lehmann, S. & Kivelä, M. Privacy and uniqueness of neighborhoods in social networks. *Sci. Rep.* **11**, 20104 (2021).
12. de Jong, R. G., van der Loo, M. P. J. & Takes, F. W. Algorithms for efficiently computing structural anonymity in complex networks. *ACM J. Exp. Algorithm.*https://doi.org/10.1145/3604908 *(2023).*
13. Sapiezynski, P., Stopczynski, A., Lassen, D. D. & Jørgensen, S. L. The Copenhagen networks study interaction data. Figshare. https://doi.org/10.6084/m9.figshare.7267433.v1. Accessed May 2022 (2019).
14. Ji, S., Mittal, P. & Beyah, R. Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey. *IEEE Commun. Surv. Tutorials* **19**, 1305–1326 (2016).
15. Li, Y. *et al.* Private graph data release: A survey. *ACM Comput. Surv.* **55**, 1–39 (2023).
16. Beigi, G. & Liu, H. A survey on privacy in social media: Identification, mitigation, and applications. *ACM Trans. Data Sci.* **1**, 1–38 (2020).
17. Sala, A., Zhao, X., Wilson, C., Zheng, H. & Zhao, B. Y. Sharing graphs using differentially private graph models. In *Proceedings of the 11th ACM SIGCOMM Internet Measurement Conference*. 81–98 (2011).
18. Proserpio, D., Goldberg, S. & McSherry, F. Calibrating data to sensitivity in private data analysis: A platform for differentially-private analysis of weighted datasets. *Proc. VLDB Endow.* **7**, 637–648. https://doi.org/10.14778/2732296.2732300 (2014).
19. Wang, Y. & Wu, X. Preserving differential privacy in degree-correlation based graph generation. *Trans. Data Privacy* **6**, 127 (2013).
20. Liu, K. & Terzi, E. Towards identity anonymization on graphs. In *Proceedings of the 4th ACM SIGMOD International Conference on Management of Data*. 93–106 (2008).
21. Zhou, B. & Pei, J. Preserving privacy in social networks against neighborhood attacks. In *Proceedings of the 24th IEEE International Conference on Data Engineering*. 506–515 (2008).
22. Hay, M., Miklau, G., Jensen, D., Weis, P. & Srivastava, S. Anonymizing social networks. In *Computer Science Department Faculty Publication Series*. Vol. 180 (2007).
23. Zou, L., Chen, L. & Özsu, M. T. K-automorphism: A general framework for privacy preserving network publication. In *Proceedings of the of the 35th VLDB Endowment*. Vol. 2. 946–957 (2009).
24. Wu, W., Xiao, Y., Wang, W., He, Z. & Wang, Z. K-symmetry model for identity anonymization in social networks. In *Proceedings of the 13th International Conference on Extending Database Technology*. 111–122 (2010).
25. Willenborg, L. & de Waal, T. *Elements of Statistical Disclosure Control* Vol. 155 (Springer, 2001).
26. Hundepool, A. *et al. Statistical Disclosure Control* Vol. 2 (Wiley, 2012).
27. González, A. & Puertas, M. L. Removing twins in graphs to break symmetries. *Mathematics* **7**, 1111 (2019).
28. McKay, B. D. & Piperno, A. Practical graph isomorphism, II. *J. Symb. Comput.* **60**, 94–112 (2014).
29. Erdős, P. & Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–60 (1960).
30. Barabási, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
31. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
32. Kunegis, J. Konect: The Koblenz network collection. In *Proceedings of the 22nd International Conference on World Wide Web*. 1343–1350 (2013).
33. Sociopatterns: Datasets. http://www.sociopatterns.org/datasets/. Accessed May 2023 (2021).
34. Rossi, R. A. & Ahmed, N. K. The network data repository with interactive graph analytics and visualization. In *AAAI* (2015).
35. Leskovec, J. & Krevl, A. *Snap Datasets: Stanford Large Network Dataset Collection*. http://snap.stanford.edu/data. Accessed May 2022 (2014).
36. Zitnik, M., Sosič, R., Maheshwari, S. & Leskovec, J. *Biosnap Datasets: Stanford Biomedical Network Dataset Collection*. http://snap.stanford.edu/biodata. Accessed May 2023 (2018).
37. Fire, M. *Data 4 Good Lab*. https://data4goodlab.github.io/MichaelFire/#section3. Accessed May 2023 (2020).
38. Aggarwal, C. C. & Yu, P. S. A general survey of privacy-preserving data mining models and algorithms. In*Privacy-Preserving Data Mining: Models and Algorithms*. 11–52 (2008).
39. Machanavajjhala, A., Kifer, D., Gehrke, J. & Venkitasubramaniam, M. l-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data (TKDD)***1**, 3-es (2007).
40. Loo, M. V. D. Topological anonymity in networks. In *Discussion Paper, Statistics Netherlands, The Hague*. https://www.cbs.nl/en-gb/background/2022/17/topological-anonymity-in-networks (2022).
41. Hagberg, A., Swart, P. & Schult, D. Exploring network structure, dynamics, and function using network. In Technical Report (Los Alamos National Lab (LANL), 2008).

Acknowledgements

This research was made possible by the Platform Digital Infrastructure SSH (http://www.pdi-ssh.nl). We would also like to thank the POPNET team (https://www.popnet.io) for various helpful suggestions and discussions.

Author contributions

R.J., M.L., and F.T. conceptualized the study and methodology. R.J. wrote the manuscript. M.L. and F.T. acquired funding, supervised and administered the project. R.J. developed the software, performed the experiments and visualized the results. All authors reviewed and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-50617-z.

Correspondence and requests for materials should be addressed to R.G.J.

Reprints and permissions information is available at www.nature.com/reprints.