Towards Statistical Disclosure Control for Complex Networks

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

[1]*Leiden University*, [2]*Statistics Netherlands*

1. INTRODUCTION

In recent decennia, research in the field of Network Science where networks are employed as a model for social and economic structures, has surged. National Statistical Institutes, especially those with access to nation-wide administrative data are in the unique position to provide relevant, accurate, and up-to-date data for the construction of population-scale networks. Indeed, Statistics Netherlands has recently constructed a 'social network' of the entire Dutch population, where nodes represent people, and edges represent real-world relations including kinship, co-workership, geographical vicinity (neighbours) and shared school [1,2]. This 'network view' on society has already yielded applications, including new ways of measuring segregation [3, 4].

The introduction of network microdata poses new challenges in the area of Statistical Disclosure Control (SDC). In contrast to traditional relational microdata where population units are represented by a set of unrelated tuples of attributes, the units represented in network data are interrelated by one or more types of edges. Although this structural information generates valuable insights, it also presents new risks of disclosure as adversaries may use (partial) knowledge of network structure to re-identify nodes or their properties.

Here, we present our ongoing research in the area of Disclosure Control for Complex Networks. Focusing on the risk of node reidentification, the central research questions revolve around attacker scenarios, how to measure and compute risk of disclosure, and approaches to alter a network to mitigate this risk.

To address these questions, we have defined and investigated the notion of *d-k*-anonymity: a graded variant of *k*-anonymity that enables one to vary the amount of structural information an adversary can use to reidentify nodes. Since computing this measure is computationally intensive, we have investigated algorithms to efficiently compute it, as well as alternative measures that can be used as approximation. Based on a wide range of computational experiments we draw conclusions on what type of knowledge makes an adversary most dangerous. We have also looked into cascading effects, *i.e.* the extent to which reidentification of one node leads to reidentification of others. Finally, we have started to explore perturbing network data to protect against reidentification.

2. APPROACH AND MAIN RESULTS

In all of the work presented here, we use networks that consist of nodes with at most a single undirected edge between each other. The nodes might represent for example people, households, or businesses. In applications these nodes are typically labelled with non-identifying attributes, such as income or health status. In our scenario, we assume the network is publicly available, and an attacker has

R.G. de Jong, M.P.J. van der Loo and F.W. Takes (2025) Towards Statistical Disclosure Control for Complex Networks, New Techniques and Technologies for Statistics, Brussels https://cros.ec.europa.eu/ntts2025.

some structural information available that can be used to narrow down the possible candidates for a target node. As an example, consider a kinship network of living persons, linked by parent-child relations. If an attacker knows, for example, that a target node has no children and a single living parent, this information can be used to narrow down the number of nodes representing the target, possibly to a single entity.

2.1. *d-k*-Anonymity

To quantify the amount of structural information an adversary might have about the network structure surrounding a node, we introduce the notion of *d-k*-anonymity [5,6]. We say that two nodes are *d-equivalent* if (1) they have the exact same neighbourhood structure up to distance *d* and (2) they occupy the same structural location in their respective neighbourhoods. A node is called *d-k anonymous*, if it is *d*-equivalent with at most *k-1* other nodes in the network. In the case of *d=1*, this models the scenario where an attacker knows how many direct neighbours a target node has, as well as all the edges between its direct neighbours. For *d=2*, this is extended to neighbours of neighbours and all edges between them, and so on.

Computing *d-k* anonymity is computationally expensive, as it requires comparing possibly many nodes, and each comparison possibly requires the computationally intensive operation of determining graph isomorphism. We have therefore designed algorithms that prevent unnecessary comparison of nodes by using heuristics to avoid isomorphism computation. For example, two neighbourhoods cannot be isomorphic if they do not have the same number of nodes and edges, which is easy to determine. The combined optimizations yield a speedup of up to four orders of magnitude as compared to the naïve approach [6].

A computational study on a wide range of model networks as well as on real networks reveals that an adversary is generally capable of reidentifying a large fraction of nodes when they have knowledge of the full neighbourhood structure of a target node, up to and including distance *d=2*. We also demonstrated that a cascading effect, where an attacker knows that a target node is linked to an already identified node, can have a further effect yielding 50% extra reidentification on average [7].

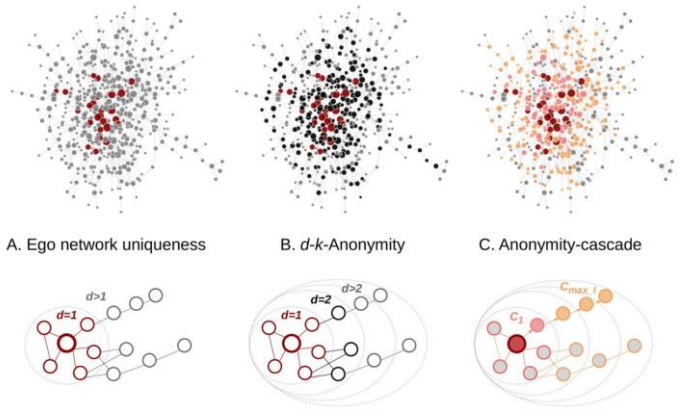

Figure 1. Measuring anonymity assuming an adversary has knowledge of node structure up to and including *d=1* (A), *d=2* (B) and accounting for a cascading

effect (C). Non-grey nodes are re-identifyable by adversary knowledge [7].

2.2. Comparing measures for k-anonymity

The *d-k-a*nonymity measure assumes that an adversary has a large amount of information available, perhaps an unrealistic amount. In that sense *d-k-a*nonymity may be a 'too strict' measure. Moreover, even with all optimisations available, computing *d-k*-anonymity can be computationally cumbersome for large networks. We have therefore explored and compared various measures of *k*-anonymity for networks. They differ from *d-k-a*nonymity in the way that equivalence between two nodes is defined.

Amongst the studied measures of equivalence are DEGREE-equivalence where nodes are equivalent when they have the same number of neighbours; COUNT equivalence, that compares the number of nodes and links the neighbourhoods of nodes; degree distribution equivalence (DEGDIST) that compares the distribution of degrees of the neighbourhoods of nodes; vertex refinement query equivalence (VRQ), that compares the degrees of all nodes in a neighbourhood; and HYBRID equivalence, that takes account of full neighbourhood structure and the degrees of the furthest nodes of the neighbourhood. We proved that these measure can be organized in a partial order according to strictness – or the amount of information they assume for the adversary [8].

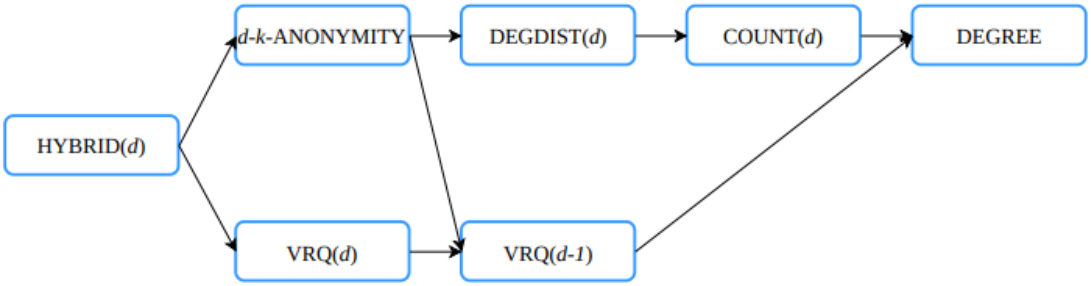

Figure 1. Strictness of anonymity measures. A → B implies A is stricter (assumes an adversary has more information) than B [8].

A computational study on model networks and a wide range of real networks demonstrated that the COUNT measure is often a good proxy for d-k anonymity while it is computationally much less demanding. Second, we see that it is generally more advantageous for an adversary to have less complete information that includes larger distances from the target node, than very complete information of the direct surroundings of a target node.

2.3. Anonymization of complex networks

As a next step in our research we have started to work on methodology to 'minimally' manipulate networks to increase the fraction of non-unique nodes [9]. In existing literature on this topic, a specific anonymization technique is often devised for a certain measure. However, we argue that as in SDC methodology for relational data, the anonymization can be treated as a separate problem from measuring anonymity. We have therefore defined several versions of the

anonymization problem that can be investigated regardless of the way anonymity ('risk of disclosure') is measured.

Variants of the anonymization problem differ in goal and boundary conditions. One may aim for full anonymization (ensuring all nodes are at least *k*-anonymous), partial anonymization (ensuring at least a fraction of nodes are at least *k*-anonymous), or one might find optimal anonymity, given a budget of perturbations. Given those variants, we use computational studies to select a measure for anonymity and found COUNT to be an overall good proxy for many applications.

We also tested a number of general perturbation techniques, including edge deletion, edge swapping and edge addition, and found that edge deletion generally requires less modifications to reach a satisfactory level of anonymity. This is due to the finding that nodes tend to me more anonymous in sparse networks where only a small fraction of all possible edges is actually realized. We used six different approaches of edge deletion that take account of the structure of surrounding nodes in varying degrees. The first and simplest method is random edge deletion (RANDOM). Second, we prioritize edges that connect two high-degree nodes (DEGMIN), or (three) a high-degree with a low-degree node (DEGDIFF). Fourth we prioritize edges that are high-impact in the sense that deleting them affects the structure of many nodes (AFF). Fifth, we prioritize edges connected to a unique node (AFF-U) or (six) nodes that affect the structure of many unique nodes (U-AFF-U).

Of all tested approaches, the algorithms that target edges that directly or indirectly affect the network structure surrounding unique nodes perform best.

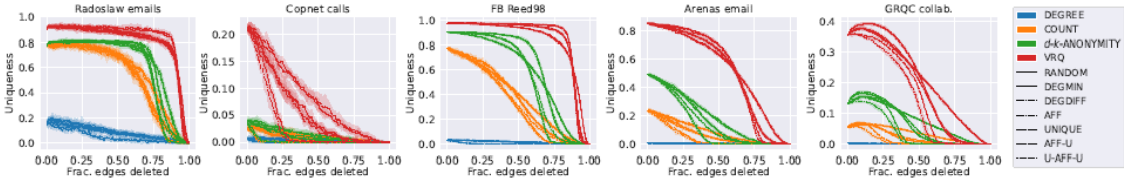

Figure 3. Fraction of unique nodes as a function of fraction of edges deleted, for various algorithms (line type) and anonymity measures (color).

3. SUMMARY AND OUTLOOK

Network science is an extremely interesting field for Statistical Institutes since it promises to investigate social and economic phenomena from the perspective of interconnected units. Moreover, Statistical Institutes are often in the unique position to construct networks from reliable administrative sources that are otherwise unavailable for (network) scientists.

In this work we focus on the problem of Statistical Disclosure Control for network data. We demonstrated, compared and categorized various ways of measuring node anonymity, and showed that (1) adversaries with knowledge of a target's surrounding network structure has a high probability of reidentifying nodes and (2) having incomplete information at a larger distance is more advantageous for an adversary then complete knowledge of the nearby structure of a target. Our research into network anonymization demonstrates that targeted edge deletion

methods that aim to affect unique nodes outperform other, simpler tested methodologies. Future work will focus on improving anonymization techniques with better targeting and allowing for explicit account of data utility.

REFERENCES

[1] Van der Laan, J., de Jonge, E., Das, M., Te Riele, S., & Emery, T. (2023). A whole population network and its application for the social sciences. European Sociological Review, 39(1), 145-160

[2] Bokányi, E., Heemskerk, E. M., & Takes, F. W. (2023). The anatomy of a population-scale social network. Scientific Reports, 13(1), 9209.

[3] Van der Laan, J., Das, M., te Riele, S., de Jonge, E., & Emery, T. (2021). Using a network of the whole population of the Netherlands to measure exposure to differing educational backgrounds. https://doi.org/10.31235/osf.io/7jtb2

[4] Kazmina, Yuliia, Heemskerk, E. M., Bokányi, E., & Takes, F. W. (2024). "Socio-economic segregation in a population-scale social network." *Social Networks* 78: 279-291

[5] Van der Loo, MPJ (2022). Topological Anonymity in Complex Networks. Technical Report, Statistics Netherlands April 21 2022 PDF

[6] R.G, de Jong, van der Loo, M.P.J., Takes, F.W. (2023). Algorithms for Efficiently Computing Structural Anonymity in Complex Networks. ACM Journal of Experimental Algorithmics 28 1—24

[7] R.G. de Jong, van der Loo M.P.J., Takes, F.W. (2024). The effect of distant connections on node anonymity. Scientific Reports 14 1156

[8] R.G. de Jong, van der Loo, M.P.J., Takes, F.W. (2024). A systematic comparison of measures for k-anonymity in networks. https://arxiv.org/abs/2407.02290

[9] R.G. de Jong, van der Loo, M.P.J., Takes, F.W. (2024). The anonymization problem in social networks. https://arxiv.org/html/2409.16163v1