UNECE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Confidentiality

1-3 December 2021, Poland

Structural uniqueness in network data.

Marieke de Vries (Statistics Netherlands)
mm.devries@cbs.nl

Abstract

A relatively new type of data that NSIs encounter is network data. For example, a network that represents family relations, neighbour relations, co-worker relations or a network that represents the connections between different businesses. This can be seen as either a new type of output or as a particular format of microdata. Hence, we need to consider the statistical disclosure part of this type of data. One of possible risk measures is related to the "structural uniqueness" in a network: when you consider a node in a network and look at its neighbourhood structure within the network, how unique is that structure? This risk measure makes use of the isomorphism definition in graph theory and can be related to the notion of k-anonymity. We will elaborate on some work that has been done in this area using unlabelled graphs and we will try to extend that work to labelled graphs.

Structural Uniqueness in Networks

Mark van der Loo*, Rachel de Jong*,**, Frank Takes**, Marieke de Vries*,
Peter-Paul de Wolf*

* Statistics Netherlands, Henri Faasdreef 312, 2492JP Den Haag.
mpj.vanderloo@cbs.nl, rg.dejong@cbs.nl, mm.devries@cbs.nl, pp.dewolf@cbs.nl

** Leiden Institute for Advanced Computer Science. University of Leiden, Niels
Bohrweg 1 2333 CA Leiden, the Netherlands. takes@liacs.nl

Abstract. Using a network representation of (economic) populations is of increasing interest to academic researchers and also at the frontiers of current research in official statistics. National Statistical institutes with access to administrative sources are in a unique position to construct societal networks that are of interest to themselves as well as to researchers. This also raises the question as to how revealing network structure is for the entities represented in it. In this paper we present a parameterized measure of k-anonymity based on the structural position of nodes in a network. We demonstrate first computational results on model networks and on the family network of the Netherlands. Moreover, we discuss how the measure can be extended to the case of labelled networks, where nodes are endowed with properties, and point out future areas of research.

1 Introduction

Over the last decades, using networks as a means to study society and economy has quickly matured into an independent scientific field, called 'social network analysis' (Newman, 2010; Barabási, 2016; Latora et al., 2017). In this area, entities like persons or companies are represented as nodes in a network and relations between them are represented as links. Important research topics include the formation of clusters, the network as a transport infrastructure, finding important nodes and pathways, and the evolution of networks. For official statisticians, network science offers a unique opportunity to study the structure of society in ways that are not possible with traditional microdata and their aggregates. National Statistical Institutes (NSIs) often have access to population scale (register) data. This unique position allows them to construct accurate networks that are hard to obtain by scholars.

For this reason, Statistics Netherlands recently created one of the first population-scale networks. In this network, each node represents a person, and there are several types of links, defined by family relations, neighbours, joint work place and joint school (van der Laan and de Jonge, 2017). The network contains more than 15 million nodes and about 39 billion edges. Such a network view of society offers unique and new ways of measuring e.g. segregation by properties such as ethnic or educational background (van der Laan et al., 2021).

Needless to say there is a large interest from researchers external to Statistics Netherlands to access such previously unavailable datasets. As these datasets contain personal data in the sense of the GDPR, this leads to the question of how network data, and results of network analyses, should be appropriately protected against disclosure. Also, one may ask the question of how insights gained from studying network data can affect disclosure risk in traditional tables and microdata.

Here, we focus on networks as microdata and we study how revealing the structural position of a node in a network can be. To illustrate this, consider the following example. Suppose that Statistics Netherlands publishes a network where nodes represent people, and links represent parent-child relations. If the nodes were labelled with the income of the corresponding persons, a user with some knowledge about the family structure surrounding a particular person, could through a structural query on the network significantly improve his estimate of that persons income or even completely reveal it when the structure is unique. Moreover, even without such labels, he might deduce some sensitive family relationships that he didn't know beforehand. Intuitively, the more a user knows about the network structure surrounding a person, the easier it is to reidentify the node.

The contribution of this paper is as follows. We introduce a parameterized measure for structural k-anonymity. We discuss some of its properties and demonstrate computational results on well-known network models and the Dutch family network. We also discuss how the current model can be extended to situations where an attacker has (partial) access to node labels in the neighbourhood of his target. We conclude with a summary and point out interesting areas for future research.

2 Measuring anonymity in complex networks

Anonymisation of nodes in networks has been studied to some extent in the Network science community. In particular, several interpretations of k-anonymity for networks have been devised. Hay et al. (2008) for example introduces a measure based on the the degree of a node and the degrees of its surrounding nodes, while Zou et al. (2009) calls two nodes k-equivalent when they occur in a k-sized orbit of the graph's automorphism group. A disadvantage of the former approach is that the existence of cycles as a revealing feature are not taken into account. In the approach of Zou et al. (2009) the complete structural position of a node in a graph is taken into account. Demanding a reasonable degree of anonymity for each node in the graph would result in an unrealistically high level of symmetry that is never observed in practical (social) networks. And although Zou's approach does allow for the most general level of protection against structural de-anonymisation, it presumes a scenario where an attacker has complete knowledge of a node's structural position in the network, which seems unlikely.

For a more complete review of literature we refer to De Jong (2021). In what follows we define a measure of k-anonymity based on the full structure surrounding a node up to and including a distance d. For example, setting $d = 1$ amounts to the assumption that an attacker knows all the information about the direct neighbours of a node. We introduce *d-k* anonymity (van der Loo, 2020), where a node is d-k-anonymous when there are $k - 1$ equivalent nodes considering their surrounding structure not further away than distance d. Intuitively this allows us to control for the expected amount of information an attacker might have about a target node as input for his structural query on the network. A more formal description follows below..

Notation 2.1. A *graph* G is a pair of finite sets (V, E) where V are *nodes* and E is a subset of $V \times V$ representing *edges* (links) between nodes. If v and w are nodes of a graph then an edge between them is denoted vw. If v is a node of G then $N(v, d)$ denotes the subgraph surrounding v up to and including distance d. Furthermore, the *degree* of a node is the number of edges it is connected to, the *distance* between two nodes is the length of the shortest path (number of edges) between them, and the *diameter* of a graph is the largest distance occurring between two nodes in the graph.

Definition 2.2 (Isomorphism)**.** Let $G = (V, E)$ and $H = (V', E')$ be two graphs. We call φ an *isomorphism* if $\varphi : V \mapsto V'$ is a bijective function such that $vw \in E$ if and only if $\varphi(v)\varphi(w) \in E'$. Two graphs G and H are called isomorphic whenever an isomorphism between them exists and this relation is denoted by $G \simeq H$.

If an isomorphism maps a graph onto itself, the function is called an automorphism. An automorphism thus intuitively permutes nodes without breaking the edges between these nodes. The set of automorphisms of G form the group $\mathrm{Aut}(G)$ which can be used to partition the nodes into equivalence classes.

Definition 2.3 (Equivalence)**.** Let S be a set. An equivalence relation on that set, denoted by $s \simeq t$ for $s, t \in S$, should satisfy the following properties:

1. $s \simeq s$ for all $s \in S$

2. if $s \simeq t$ then $t \simeq s$ for all $s, t \in S$

3. if $s \simeq t$ and $t \simeq u$ then $s \simeq u$ for all $s, t, u \in S$.

Definition 2.4 (d-k-anonymity)**.** Let $G = (V, E)$ and $d \in \mathbb{N}$. We say that v and w in V are d-equivalent when

1. $N(v, d) \simeq N(w, d)$; and

2. There is an isomorphism $\varphi : N(v,d) \mapsto N(w,d)$ such that $\varphi(v) = w$.

We then write $v \simeq_d w$. We denote with $[v]_d$ all nodes in G that are equivalent to v. We say that a node is d-k-anonymous when the size of this equivalence class $[v]_d$ equals k.

The first demand states that two nodes need to have 'indistinguishable' environments (i.e. isomorphic neighbourhoods). The second demand states that they need to play the same role in their respective environments. The following properties of d-k anonymity are not hard to prove. First, one can demonstrate by using general properties of graph isomorphisms that $\simeq_d$ indeed satisfies all properties of an equivalence relation. Second, we have the desirable property that d-k anonymity is non-increasing as a function of d. In other words, denoting with $a(v,d)$ the d-k anonymity of v, we have $a(v,d+1) \leq a(v,d)$. This can be interpreted as the intuitive fact that when an attacker knows more about a node's surroundings, it will not become harder to find that node in the network and in practice will often become easier.

3 Computational results

To compute structural anonymity it is necessary to compare (possibly many) subgraphs for isomorphism. Such calculations are typically very computationally intensive, and we have therefore implemented a few ideas to make these computations more tractable. The first idea is to use the fact that two nodes can be $d+1$ equivalent only when they are d-equivalent. So once we determine an equivalence class $[v]_d \subseteq V$, we only need to compare nodes within that class to test whether they are equivalent at $d+1$. Starting with $[v]_0 = V$, we can thus recursively compute all sets of equivalence classes up to any d. Computational effort is then determined by two competing forces: on one hand, as d increases, larger subgraphs need to be tested for isomorphism. On the other hand, the number of comparisons within each class will typically decrease. The second idea is to avoid graph isomorphism calculations by rejecting candidates that have no chance of being isomorphic. This can be done by comparing graph invariants that are easily computed, such as the number of nodes and the number of edges in two graphs. If such graph invariants are not equal to each other for two graphs, then they can not be isomorphic. De Jong (2021) implemented several versions of the anonymity measure using the `nauty` library of McKay and Piperno (2014) for isomorphism computations.

Figure 1 summarizes some computational results on widely used network models. In each figure, each dot represents a node in a graph. The vertical axis labels the anonymity while the horizontal axis labels the edge density (average number of edges per node) of graphs generated under the respective models.

The first model is the Erdős-Rényi (ER) graph (Erdos and Rényi, 1960). In this model a graph is generated by starting with n nodes, and each $n(n-1)/2$ node pair is linked with probability p. This is the simplest random network model that serves as a benchmark for many methods in network science. The second model is the Barabási-Albert (BA) model (Barabási and Albert, 1999). Here, graphs are generated by adding nodes with a fixed number of dangling links, that attach with higher probability to nodes that already have many links. It serves as an important model since the power-law degree distribution resulting from this mechanism is encountered in many networks found in practice. The third model is a power-law cluster (PC) graph (Holme and Kim, 2002). This is similar to a Barabási-Albert graph, except that there is a second probability p_2 that controls whether a triangle is closed when a node is added. The number of triangles in a graph is a measure for the amount of clustering taking place, and higher p_2 implies more clustering, which is often observed in real-world networks.

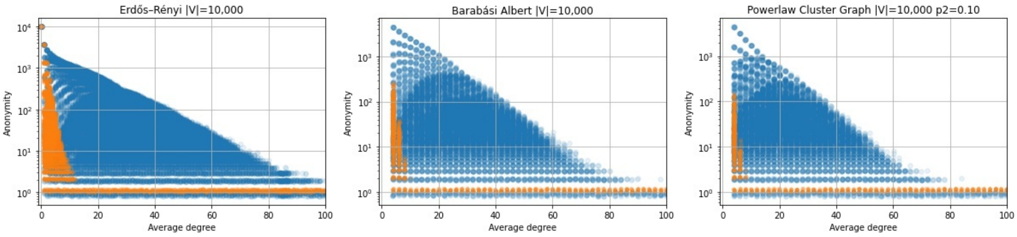


Figure 1: Anonymity values for $d = 1$ (blue) and $d = 2$ (orange) for nodes in Erdős-Rényi (left) Barabási-Albert (middle) and Powerlaw-cluster graphs (right). The horizontal axis represents the average number of edges per node. Each figure shows results of ten numerical experiments. Reproduced from De Jong (2021).

From Figure 1 two trends are immediately visible. First, in all models the anonymity decreases with higher edge density. As the number of edges increases, more variation in the structure of subgraphs becomes possible so this is not unexpected. We do see that the decrease in anonymity as a function of edge density is stronger in BA and PC networks than in ER graphs. It seems easier to be anonymous in a completely random ER network than in networks with a certain (probabilistic) structure enforced. Second, we see that anonymity drops sharply going from $d = 1$ to $d = 2$. Especially for the studied BA and PC networks it is true that knowing the neighbourhood of a node up to two steps away will allow the attacker to uniquely identify all nodes in nearly all graphs. This conclusion is supported by a theoretical result of Hay et al. (2008), who demonstrates that in dense ER graphs, the expected size of $d = 2$ equivalence classes goes to zero as $|V| \to \infty$ (but note that in Hay et al. (2008) equivalence is weaker compared to the measure we use here; in their case only the degree structure surrounding a node is taken into account).

We also computed anonymity of nodes in the Dutch family network, for $d = 1, 2, \ldots, 5$. This network consists of 15.7 milion nodes linked by parent-child relations. As an illustration, Figure 2 shows the second-largest connected component of the network, where nodes are colored according to having high (pink) to low (gray) anonymity. Again, we see that as an adversary learns more about a person's family relations, anonymity drops sharply. In Table 1 we summarize how many nodes have an anonymity below threshold values of $k = 1$, $k = 2$, and up to $k = 5$ or more for different values of d. If we take $k = 3$ (an often recommended value for disclosure control of microdata), we see that at $d = 1$ three out of more than 15 million are 3-anonymous and thus have a probability of 1/3 of being correctly identified. For $d = 2$ more than 5600 nodes have that probability. At $d = 5$, almost 600 thousand nodes have a probability 1/3 of being re-identified.

The calculations on the model networks and on the family demonstrate that for network data, (partial) knowledge of the structural position of a node in a network can significantly increase the risk of re-identification or disclosure of a property.

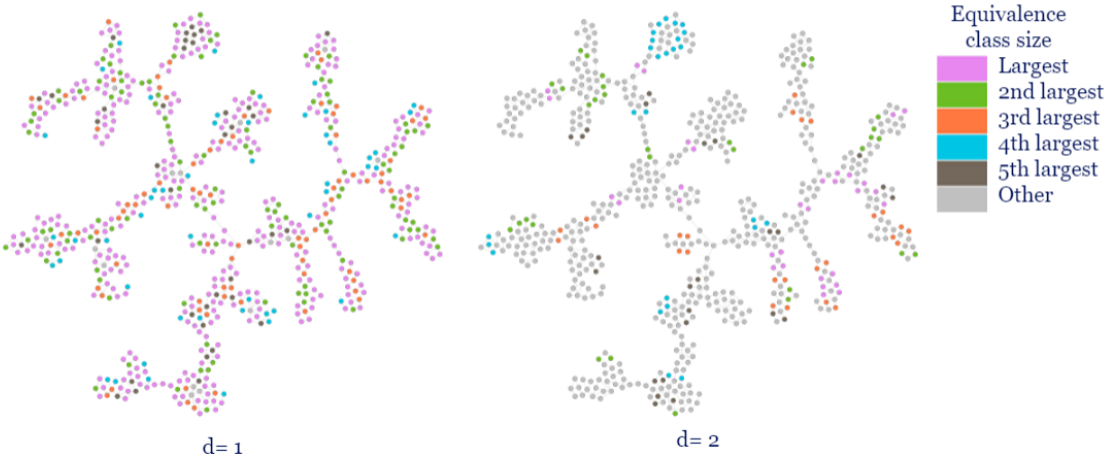


Figure 2: The second-largest component of the Dutch family network. Nodes represent persons, links represent an undirected parent-child relationship. Nodes are colored according to their equivalence class. Reproduced from De Jong (2021).

4 Extension to labelled graphs

While the addition of graph labels enhances the possibilities for research on the network, they also increases the challenge for statistical disclosure control. When protecting 'ordinary' microdata, we will check for properties like k-anonymity on a set of quasi-identifiers in the dataset, meaning any combination of quasi-identifiers needs to be occurring least k (or zero) times (Sweeney, 2002). Protecting the network however requires more work than if we were to protect the microdata that is published without its underlying structure.

We need to account for network structure in two different ways. First, the structural position of a node in a network is another characteristic of a person

k	d=1	d=2	d=3	d=4	d=5
1	7	17,383	335,143	1,423,905	2,728,574
2	4	12,784	179,428	714,844	1,369,664
3	3	5,637	87,150	374,739	581,271
4	4	6,071	84,660	312,008	428,872
5	10	4,130	47,440	119,530	136,560
≥ 5	15,760,693	15,714,715	15,026,900	12,815,695	10,515,780

Table 1: Number of $\leq k$ anonymous vertices in full family network. Reproduced from De Jong (2021).

in our network, and leads to increased risk in its disclosure. Furthermore, not only do the labels increase the information known about a node, they also give more information about the nodes in its neighbourhood. For example, knowing a node is related to a person with label ‘female’ is more disclosive than just knowing it is related to another node. A measure like the one proposed in section 2 could work well for taking this into account. Indeed, in this section we will extend the definition of *d*-*k* anonymity to labelled graphs.

Notation 4.1. We define a labelled graph G by the 6-tuple $G = (V, E, L_V, L_E, l_v, l_e)$. Here V, E are the vertex and edge set respectively, L_V, L_E the set of labels on vertices and edges respectively, and $l_v : V \mapsto L_V$ and $l_e : E \mapsto L_E$ functions that assign labels to the vertices and edges. That is, for $u, v \in V$ and $uv \in E$, we have $l_v(u) = \alpha$ for some $\alpha \in L_V$ and $l_e(uv) = \beta$ for some $\beta \in L_E$.

Isomorphism for labelled graphs, where a labelled graph is defined as a graph such that vertices and edges are assigned labels, is defined in (Hsieh et al., 2006) as follows:

Definition 4.2 (Isomorphism for labelled graphs)**.** Graphs $G = (V, E, L_V, L_E, l_v, l_e)$ and $G' = (V', E', L'_V, L'_E, l'_v, l'_e)$ are *isomorphic*, written $G \simeq H$, when there exists a bijective function $\varphi : V \mapsto V'$ such that

1. For all $u \in V$ we have $l_v(u) = l'_v(\varphi(u))$,
2. For all $u, v \in V$ we have $uv \in E \iff \varphi(u)\varphi(v) \in E'$, and
3. For all $uv \in E$, we have $l_e(uv) = l'_e(\varphi(u)\varphi(v))$.

This is similar to the isomorphism defined in definition 2.2, but with an added requirement of preserving the labels on both edges and vertices. This definition of isomorphism also preserves the adjacency relations between labelled vertices, as can be seen in Example 6.1 (see Appendix).

Let $G = (V, E, L_V, L_E, l_v, l_e)$ be a labelled graph and $Aut(G)$ the group of automorphisms, where an automorphism on a labelled graph is defined as an isomorphism on a labelled graph that maps it onto itself.

Definition 4.3 (d-k-anonymity for labelled graphs)**.** Let $G = (V, E, L_V, L_E, l_v, l_e)$ and $d \in \mathbb{N}$. We say that v and w in V are d-equivalent, if

1. $N(v, d) \simeq N(w, d)$; and
2. There is an isomorphism $\varphi : N(v, d) \mapsto N(w, d)$ such that $\varphi(v) = w$.

We then write $v \simeq_d w$. Again we use $[v]_d$ to denote all nodes in G that are d-equivalent to v, and say v is d-k-anonymous if the size of the equivalence class $[v]_d$ equals k.

While structural uniqueness of nodes for unlabelled graphs following from definition 2.4 is preserved by definition 4.3, we naturally expect the uniqueness of the nodes to grow by adding more information to them.

Theorem 4.4. *Let $G = (V, E, L_V, L_E, l_v, l_e)$ be a graph, $Aut(G)$ be the group of automorphisms of G and $Aut(G')$ be the group of automorphisms on the graph $G' = (V, E)$, the graph G where the labels on the edges and vertices are removed. Then $Aut(G) \subseteq Aut(G')$.*

Proof. Let $\phi \in \mathrm{Aut}(G)$. We will prove $\phi \in \mathrm{Aut}(G')$. By definition, we have that $\phi : V \mapsto V$ is a bijection such that for all $u, v \in V$, we have $uv \in E$ if and only if $\phi(u)\phi(v) \in E$. Thus, the isomorphism ϕ on G is also an isomorphism on G'. Therefore, $\phi \in \mathrm{Aut}(G')$ and thus $\mathrm{Aut}(G) \subseteq \mathrm{Aut}(G')$. □

Corollary 4.4.1. *Equivalent classes for labelled graphs are at most of the same size as their non-labelled counterparts. Therefore, the uniqueness of nodes and thus the risk of disclosure is not lower (and in practice often higher).*

See also Example 6.3 in the appendix. As expected, the more information we add to the network, the more risk of disclosure we have.

5 Conclusion and outlook

In recent years, studying society in terms of complex networks has become a vivid scientific field. Official statisticians in many countries have a unique position that enables them to construct population-scale networks that can be used to derive new statistical products, including making anonymized network data available for scientific research.

In this paper we have demonstrated a new quantity for measuring the anonymity of nodes in a network. The measure is based on comparing the full structural position of nodes in their (immediate) neighbourhood and is parametrized by the size of the neighbourhood that is taken into account. The measure therefore quantifies the level of anonymity against an adversary that has (partial) knowledge of the network surroundings of a node. We have shown that it is feasible to compute this measure on large scale networks such as the Dutch family network with more than 15 million nodes. The computational results obtained thus far relate to unlabelled graphs, that is: networks where nodes and edges have no properties apart from their structural position. Researchers are often interested in properties of the nodes and/or edges as well, hence we also propose a generalisation to labelled graphs.

Our theoretical and computational results demonstrate that structural information can in principle be very revealing. Knowledge about family relations up to and including the second degree, even without taking account of parent-child directionality, de-anonymizes thousands of nodes in the family network.

The current work is still in its infancy and there are many open questions, both of methodological and practical nature. On the methodological side, extensions to labelled or multi-layered networks are obvious and interesting generalisations. There is also a need for suitable anonymisation techniques, based on for example label suppressions, or edge perturbations.

On the practical side we need to realize that the current work is based on a scenario where an adversary uses partial (yet perfect) knowledge of the network surrounding a node to re-identify a node or disclose a property from network data. For a full assessment of disclosure risk we would also need to asses how likely it is that an adversary actually obtains such information.

Moreover, we have thus far treated network data as a form of 'microdata'. We currently have little means to assess disclosure risk imposed by network-based statistics, or the disclosure risk arising from combining network data with other publications. With increased interest in network science as a means to model society, these questions will sooner rather than later need to be addressed by the official statistics community.

References

Barabási, A.-L. (2016). *Network Science.* Cambridge, UK: Cambridge University Press.

Barabási, A.-L. and R. Albert (1999). Emergence of scaling in random networks. *science 286*(5439), 509–512.

De Jong, R. G. (2021). Measuring structural anonymity in complex networks. Master's thesis, Leiden University, The Netherlands.

Erdos, P. and A. Rényi (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci 5*(1), 17–60.

Hay, M., G. Miklau, D. Jensen, D. Towsley, and P. Weis (2008). Resisting structural re-identification in anonymized social networks. *Proceedings of the VLDB Endowment 1*(1), 102–114.

Holme, P. and B. J. Kim (2002). Growing scale-free networks with tunable clustering. *Physical review. E, Statistical, nonlinear, and soft matter physics 65*(2 Pt 2), 026107.

Hsieh, S.-M., C.-C. Hsu, and L.-F. Hsu (2006). Efficient method to perform isomorphism testing of labeled graphs. In M. L. Gavrilova, O. Gervasi, V. Kumar, C. J. K. Tan, D. Taniar, A. Laganá, Y. Mun, and H. Choo (Eds.), *Computational Science and Its Applications - ICCSA 2006*, Berlin, Heidelberg, pp. 422–431. Springer Berlin Heidelberg.

Latora, V., V. Nicosia, and G. Russo (2017). *Complex Networks: Principles, Methods and Applications.* Cambridge, UK: Cambridge University Press.

McKay, B. D. and A. Piperno (2014). Practical graph isomorphism, ii. *Journal of symbolic computation 60*, 94–112.

Newman, M. (2010). *Networks: An introduction.* Oxford, UK: Oxford University Press.

Sweeney, L. (2002). k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10*(05), 557–570.

van der Laan, J., M. Das, S. te Riele, E. de Jonge, and T. Emery (2021). Measuring educational segregation using a whole population network of the netherlands. https://osf.io/preprints/socarxiv/7jtb2/.

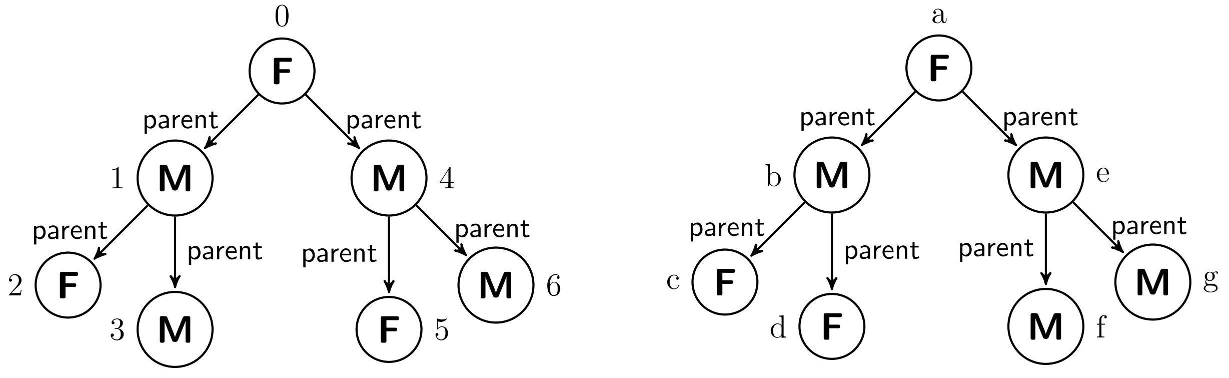
van der Laan, J. and E. de Jonge (2017). Producing official statistics from network data. In *The 6th international conference on networks and their applications*, Lyon, France, pp. 288.

van der Loo, M. P. J. (2020). Topological anonymity in networks. Technical report, Statistics Netherlands. Internal report.

Zou, L., L. Chen, and M. T. Özsu (2009). K-automorphism: A general framework for privacy preserving network publication. *Proceedings of the VLDB Endowment 2*(1), 946–957.

6 Appendix

Example 6.1. Let G and H be two graphs as given below, representing some familial relationship



Graph G

Graph H

The vertices are labelled with gender, where M represents male and F represents female. The edges contain the relationship ‘is parent of’. If all labels were removed, it is clear that $G \simeq H$.

In both graphs we have three women and four men, so we could find a bijection on the vertex sets of G and H that preserves the labels of the vertices. One such function $\phi : \{0,1,2,3,4,5,6\} \mapsto \{a,b,c,d,e,f,g\}$ maps the vertices as follows:

$$\phi = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ a & b & c & f & e & d & g \end{pmatrix}$$

The labels of the vertices indeed match: the women in G, $(\{0,2,5\})$, are mapped to women in H, $(\{a,c,d\})$, and likewise all men in G are mapped to men in H. So, for all vertices in our graph G, the label of the vertex is equal to the label of the mapped vertex. However, once we start checking the edges we run into issues. If we were to check the labelling on the edges, we find that the edges $\{01,04,12,13,45,46\}$ in G should be mapped to $\{ab,ae,bc,bf,ed,eg\}$ in H - except, both edges bf and ed are not in H.

There are two bijections on the vertices that preserve the edges, namely

$$\psi = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ a & b & c & d & e & f & g \end{pmatrix} \quad \text{and} \quad \rho = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ a & e & f & g & b & c & d \end{pmatrix},$$

but for both the labels are not preserved: $l(3) \neq l'(\psi(3)) = l'(d)$ and $l(2) \neq l'(\rho(2)) = l'(f)$.

It is impossible to find an isomorphism on the labelled graphs G and H, even though the structure of the graphs without labels is isomorphic and the number of ‘male’ and ‘female’ labels on both graphs is equal.

While it might be obvious that no single father with two daughters can be mapped to a single father with a daughter and a son, more importantly the example shows that even if two grandmothers both have two sons with both two children, of which two grandsons and two granddaughters, they cannot be mapped to each other by an isomorphism if the distribution of those grandchildren over the sons is different.

Lemma 6.2. Let $G = (V, E, L_V, L_E, l_v, l_e)$ and $G' = (V', E', L'_V, L'_E, l'_v, l'_e)$ be isomorphic, $\varphi : V \mapsto V'$ be an isomorphism and $j \in \mathbb{N}$. Let $v \in V$. Then it holds that $N(v, j) \simeq N(\varphi(v), j)$.

Proof: As φ is an isomorphism mapping V to V', we know that:

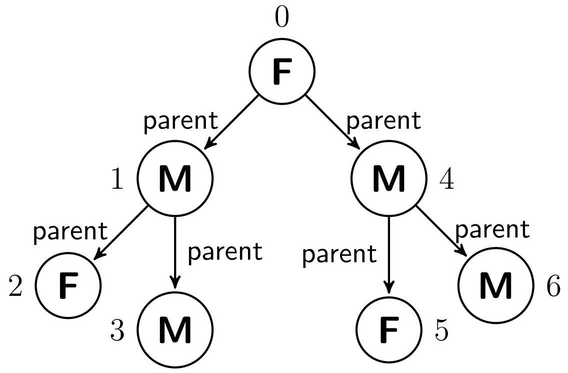
- For all $u \in V$ we have $l_v(u) = l'_v(\varphi(u))$,
- For all $u, v \in V$ we have $uv \in E \iff \varphi(u)\varphi(v) \in E'$, and
- For all $uv \in E$, we have $l_e(uv) = l'_e(\varphi(u)\varphi(v))$.

Recall definition 4.2: for two labelled graphs to be isomorphic, there needs to be an bijective function that fulfils the three requirements. Let $\varphi : V \mapsto V'$ be an isomorphism, $N(v, j)$ be the j-th order neighbourhood for a $v \in V$, and $N(\varphi(v), j)$ the j-th order neighbourhood for $\varphi(v) \in V'$. Then

1. Since φ is an isomorphism, we have if uv is an edge in $N(v, j) \subseteq G$ then $\varphi(u)\varphi(v) \in E'$. Furthermore, we have that $\varphi(u)\varphi(v) \in N(\varphi(v), j)$: since distance between nodes is graph invariant meaning it does not change under isomorphisms, we have that $d(v, u) \leq j$ implies $d(\varphi(u), \varphi(v)) \leq j$.
2. Since $l_v(u) = l'_v(\varphi(u))$ for all $u \in V$, this holds for all $u \in V(N(v, j)) \subseteq V$.
3. Similarly, the labels on the edges are preserved.

Thus indeed $N(v, j) \simeq N(\varphi(v), j)$.

Example 6.3. Recall the graphs in example 6.1. We can look at the structural uniqueness of one of these graphs G.



The diameter of this graph is 4, so we can check the d-equivalency of the nodes $v \in V$ up until $d = 4$, and compare it to the topological anonymity of the nodes in the unlabelled graph:

d v	0	1	2	3	4
0	3	1	1	1	1
1	4	2	2	2	2
2	3	2	2	2	2
3	4	2	2	2	2
4	4	2	2	2	2
5	3	2	2	2	2
6	4	2	2	2	2

d v	0	1	2	3	4
0	7	1	1	1	1
1	7	2	2	2	2
2	7	4	4	4	4
3	7	4	4	4	4
4	7	2	2	2	2
5	7	4	4	4	4
6	7	4	4	4	4

Table 2: Size of the equivalence classes $[v]_d$ for the labelled (left) and unlabelled (right) graph.

All nodes have equal or bigger equivalence classes $[v]_d$ in the unlabelled case compared to the labelled graph. This means that the labelled graph poses more problems for statistical disclosure control than its unlabelled counterpart.