Discussion Paper

# Topological anonymity in networks

Mark van der Loo

**April 21, 2022**

**Network data is of increasing interest for official statisticians. Publishing such data introduces re-identification risks that differ from tabular data or microdata records, since the network structure may yield revealing clues about the identity of a node. In this work we develop a measure of node-anonymity that is based purely on the network structure surrounding a node. We show that our definition has some desirable properties and we evaluate this measure on a small scale-free network. We point out some avenues for improving upon the current computational complexity of the algorithms implementing the anonymity measure, and discuss the most important avenues for future research.**

# Contents


1 Introduction 4

2 Scenario 5

3 Anonymity 6
- 3.1 Notation and concepts 7
- 3.2 Topological anonymity 9
- 3.3 Anonymity for general neighbourhoods 11

4 Algorithms 13

5 An example 14

6 Discussion and possible ways forward 18

7 Summary and conclusion 19

References 20

A Proofs 23
- A.1 Equivalence relation for nodes 23
- A.2 Proof of Theorem 10 23
- A.3 Proof of Theorem 14 24

## Note

This public technical report was published as an internal research report on 26 October 2020 at Statistics Netherlands by the author. The Master's thesis of de Jong (2021) built on this report, implementing amongst others a concept algorithm that improves the perfomance of the implementation mentioned in this paper with factors up to $10^4$. The work was has also been presented at two conferences (de Jong et al., 2021a,b). We publish this technical report so the original theoretical work can be referenced.

The author was unaware of existing literature on Network Anonymity, until after completing the original report. Notably, the concept of 'structural anonymity' (which we call topological anonymity) has been described in a technical report of Hay et al. (2007). Zou et al. (2009) describes a strict type of anonymity (called $k$-automorphism anonymity) which in one limit coincides with the definition developed here. An recent overview of developments in this area is also given by Ji et al. (2016).

# 1 Introduction

In recent years, Statistics Netherlands has been researching the prospects of Complexity Science (Newman, 2011), for official statistics, and of that of Network Science in particular. Indeed, van der Laan and de Jonge (2017, 2019) have demonstrated that when existing data is studied from the perspective of network science, interesting new insights and statistical products may be developed.

At the same time, network researchers outside of Statistics Netherlands are highly interested in working with networks derived from microdata owned by Statistics Netherlands. Sharing network data, for example in the form of Scientific or even Public use files raises the issue of Statistical Disclosure Control. Statistical offices, and Statistics Netherlands in particular are prolific in protecting statistical units represented in tabular or relational data against disclosure (Hundepool et al., 2012; Willenborg and De Waal, 2001). Network data fundamentally differs from these data types because its information content is defined by the way nodes are connected by edges. Such a topology ('connectedness') is less relevant for relational or tabular data, although network models do play a role in certain table-protection methods. This means that an adversary that wishes to re-identify a node in a network has an extra tool at his disposal: the network structure surrounding a node.

In Statistical Disclosure Control theory for relational microdata, the risk of disclosure is based on the concept of *anonymity*. This is a quantitative measure that measures the non-uniqueness of an individual in a population or data set, relative to its attributes. An often-used measure of anonymity for relational data sets was defined formally by Samarati and Sweeney (1998); Samarati (2001). This definition can be restated

informally as 'an individual represented by a record in a relational data set is said to be $k$-anonymous when there are $k-1$ individuals like it[1)]'.

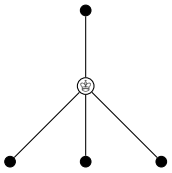
In this paper we develop a mathematical definition of anonymity for nodes in a network, demonstrate some of its properties, and present a few first computational results. We will also outline the most important open questions, and directions for future research.

# 2 Scenario

In order to assess the risk of disclosure it is necessary to identify a scenario that defines the information released by a statistical institute and the information that an adversary may have at his disposal. Since the focus of this paper is on developing concepts and theory, we shall focus on a minimal scenario that can serve as a baseline case.

For our scenario we shall assume that a statistical institute wishes to publish a network with labeled nodes. For example, the nodes may represent persons or businesses, and links may represent a family relation, or a supplier-client relationship. Nodes are labeled with a property, say income or turnover. We shall be interested in how easy or difficult it is for an adversary to re-identify a certain node by using externally gained knowledge about the network structure surrounding that node.

For an example, suppose that the nodes are people living in the Netherlands, and that the links are (undirected) parent-child relations. Suppose further that the adversary wishes to (partially) re-identify a certain person in the network. Say, the king of the Netherlands. Since it is public knowledge that he has three children, and one living parent, our adversary can query the network for nodes that have a network neighbourhood similar to the king's:



(1)

where we labeled the sought-after node with an identifying symbol for visual clarification (but this is not the case in the full network). Now, if the network contains $n$ nodes then without any further information, the probability of re-identifying the king (or any node, for that matter) out of $n = 17$ million nodes equals $1/n \approx 5.9 \times 10^{-8}$. In 2017 there were about 2 million parents with 3 children[2)]. Suppose that half of them have a living parent, this means that the number of candidate nodes decreases to about 1 Million, increasing the disclosure risk to about $10^{-6}$: an increase of two orders of magnitude.

Continuing this example, we can imagine that our adversary uses extra information about the king's network surroundings, for example that he has a sibling who also has

[1)] Samarati's original definition applies to whole data sets, and assigns to a data set the anonymity of the least anonymous individual.

[2)] https://www.cbs.nl/en-gb/news/2017/19/one-in-a-hundred-mothers-have-more-than-five-children

three children. Hence, the search can be narrowed down further using a more extensive search pattern, thereby increasing the disclosure probability.

From this example, two elements emerge that define the risk of disclosure. On one hand there is the *disclosure probability*, which is conditional on the information an adversary has. On the other hand there is the likelihood that this information actually will be obtained by an adversary. Indeed we may express the disclosure risk $r$ for a certain node $v$ in the network as

$$r(v, I) = p(v|I)p(I) \tag{2}$$

where $p(v|I)$ is the disclosure probability, given certain information $I$, and $p(I)$ the probability that an adversary will have this information. Observe that in this formulation one may always assume the worst case scenario by setting $p(I) = 1$. Generalizing a little bit, we see that whichever way $p(v|I)$ is determined, it is desirable that when the amount of information reflected by $I$ increases, $p(v|I)$ should increase as well. We would furthermore like $p(v|I) = 1/n$ when the adversary has no information, except that $v$ is one of the $n$ nodes in a network. For $p(I)$ on the other hand, one would expect it to decrease when $I$ increases, as it would cost an adversary more effort to collect more information.

The remainder of this paper is devoted to quantifying $p(v|I)$. Regarding the information $I$, we will focus on network structure surrounding the node. Thus, we ignore node properties such as the income in the previous example; leaving the integration of these properties for future work.

Finally, we note that this work is in no means finished. This paper is aimed as a progress report, showing our current knowledge and open questions. It is mainly aimed as a reference for further research.

# 3 Anonymity

The disclosure probability of an individual in a relational data set is inversely related to its *anonymity*: the number of individuals that share a set of attribute values. For networks, we shall in a similar fashion quantify anonymity of a node $v$ as the number of nodes that have the same surrounding network structure. One may choose the 'amount' of surroundings that are taken into account by considering only nodes that are no more than $j$ steps removed from our original node $v$.

This allows us to define anonymity $a$ as a function $a(v, j)$ that assigns the 'number of equivalent nodes' to a node $v$. In a way to be made precise in Section 3.2, two nodes are 'equivalent' when they have similar surroundings, and play the same role in their respective surroundings. It will be shown that distance $j$ can be interpreted as an amount of information $I$ as in Equation (2). In the sense that the disclosure probability $p(v|j) = a(v, j)^{-1}$ decreases as a function of $j$, and that $p(v|0) = 1/n$ ($n$ the number of nodes) as desired. Furthermore, when $j$ equals the diameter of the network, the disclosure probability of a node is equal to the size of its *orbit* in the network. This is a natural upper limit that is induced by the symmetries of the network.

## 3.1 Notation and concepts

We introduce the technical notations and definitions used in this paper. The terms *network* or *graph* refer to a graph that may or may not be connected, and may or may not be directed. There may, or may not be multiple edges between each pair of nodes and self-loops are not excluded. Graphs are denoted $G$, $H$, or $F$, their node and edge sets are denoted respectively $V$ and $E$, or $V(G)$ and $E(G)$ for disambiguation. Nodes are denoted $v$, $w$, or $u$, and an edge between $u$ and $v$ is denoted $uv$. The *degree* of a node is the number of edges it is connected to. The (proper) subgraph relation is denoted $\subset$ ($\subseteq$). The *distance* between two nodes is denoted $d(v, w)$ and is defined as the length of the shortest path between them (see e.g. Diestel (2000)). Additionally we assume $d(v, w) = \infty$ if there is no path from $v$ to $w$. The *components* of a graph are subgraphs that have no connections between them. The diameter $\mathrm{diam}(G)$ of a graph is the largest distance found between any two nodes of $G$. Finally, let $U$ be a subset of $V$. The subgraph *induced* by $U$ is defined by the node set $U$ and all edges in $G$ between the nodes in $U$.

The concepts of *neighbourhood*, *isomorphism*, and *automorphism* play an important role throughout the paper, and are therefore introduced formally.

**Definition 1.** Let $G$ be a graph, $v$ one of its nodes and $j$ a non-negative integer or $\infty$. The *$j$th order neighbourhood* $N(v, j)$ of a node $v$ is the subgraph of $G$ induced by all nodes with distance $d(v, w) \leq j$.

In some literature this is referred to as the *closed neighbourhood* of $v$ because it includes $v$. Observe that for all nodes $v$ of $G$ we have $N(v, \mathrm{diam}(G)) = G$, $N(v, 0) = (\{v\}, \{\})$, and $N(v, j) \subseteq N(v, j + 1)$.

**Definition 2.** Given two graphs $G$ and $H$. An *isomorphism* is a bijection $\phi : V(G) \mapsto V(H)$ such that if $vw \in E(G)$ then $\phi(v)\phi(w) \in E(H)$.

Two graphs $G$ and $H$ are called *isomporhic*, denoted $G \simeq H$ when, there is at least one isomorphism between them.

Intuitively, isomorphisms are maps between graphs that preserve structure. A *graph invariant* is a property of a graph that does not change under isomorphisms. Examples of graph invariants include the graph diameter, and the distance between any two nodes. In particular, this leads to the observation that the image of a $j$th order neighbourhood of a node $v$ is an isomorphic $j$th order neighbourhood of the image of $v$.

**Observation 3.** Given two isomorphic graphs $G$ and $H$, let $\phi : V(G) \to V(H)$ be an isomorphism, and let $j$ be a nonnegative integer or $\infty$. We have

$$G \supseteq N(v, j) \simeq N(\phi(v), j) \subseteq H, \tag{3}$$

for all nodes $v \in V(G)$.

*Proof.* Since distances are a graph invariant, we have that if $v'$ is in $N(v, j)$, then $\phi(v')$ is in $N(\phi(v), j)$. Furthermore, by definition of isomorphism we have that if $uw \in E(N(v, j))$ then $\phi(u)\phi(w) \in E(N(\phi(v), j))$. □

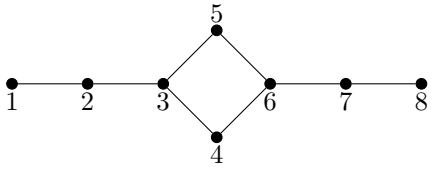
More generally, one sees that if $G \simeq H$ and $G' \subseteq G$, then the image $H'$ of $G'$ under an isomorphism connecting $G$ and $H$ is isomorphic to $G'$. In other words, *having a certain subgraph* is a graph invariant.

**Definition 4.** An *automorphism* is an isomorphism of a graph onto itself.

Automorphisms permute nodes in a graph without breaking links. They are traditionally denoted with $\pi$. The set of automorphisms of a graph is denoted $\mathrm{Aut}(G)$ and it has the structure of a *group* under function composition[3)]. The group of automorphisms partitions the nodes into *equivalence classes* of nodes that are permuted into each other under an automorphism of the graph. The group-theoretical term for the equivalence class of a node induced in this way is called its *orbit*, and the technical definition can be written as

$$\mathrm{Orbit}(v) = \bigcup_{\pi \in \mathrm{Aut}(G)} \{\pi(v)\}. \tag{4}$$

**Example 5.** Consider the following graph $G$.




We have $\mathrm{diam}(G) = 6$, and the automorphism group $\mathrm{Aut}(G)$ consists of four node permutations $e, a, b, c$ that can be written as follows (using Cauchy's notation for permutations).

$$e = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{pmatrix}, \; a = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 8 & 7 & 6 & 4 & 5 & 3 & 2 & 1 \end{pmatrix}$$

$$b = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 1 & 2 & 3 & 5 & 4 & 6 & 7 & 8 \end{pmatrix}, \; c = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 \end{pmatrix}$$

The multiplication rules are given by $a^2 = b^2 = c^2 = e$ and $ab = ba = c$ (incidently, this shows that $\mathrm{Aut}(G)$ is isomorphic to Klein's four-group $V_4$). We may interpret each element of $\mathrm{Aut}(G)$ as a function. For example $e(6) = 6$ and $a(2) = 7$. Using Equation (4), we can compute the orbit of each element.

$$\mathrm{Orbit}(1) = \{e(1)\} \cup \{a(1)\} \cup \{b(1)\} \cup \{c(1)\} = \{1, 8\} = \mathrm{Orbit}(8)$$
$$\mathrm{Orbit}(2) = \{e(2)\} \cup \{a(2)\} \cup \{b(2)\} \cup \{c(2)\} = \{2, 7\} = \mathrm{Orbit}(7)$$
$$\mathrm{Orbit}(3) = \{e(3)\} \cup \{a(3)\} \cup \{b(3)\} \cup \{c(3)\} = \{3, 6\} = \mathrm{Orbit}(6)$$
$$\mathrm{Orbit}(4) = \{e(4)\} \cup \{a(4)\} \cup \{b(4)\} \cup \{c(4)\} = \{4, 5\} = \mathrm{Orbit}(5).$$

[3)] This means that if $\pi$ is an automorphism of $G$, then so is its inverse $\pi^{-1}$; when $\pi$ and $\pi'$ are automorphism of $G$, then so is $\pi\pi'$ (applying $\pi$ after applying $\pi'$); there is a unique unit element $e$ of $\mathrm{Aut}(G)$ such that $e\pi = \pi e = \pi$ for all $\pi$.

## 3.2 Topological anonymity

Anonymity is a measure of non-uniqueness. To measure non-uniqueness we seek a way to partition the nodes in a graph into classes of nodes that are in some sense equivalent. One way to go about this, and this is the way we will do it here, is to find an appropriate formal *equivalence relation* between nodes. This then immediately yields disjunct equivalence classes. The size of an equivalence class is then a measure of anonymity.

We remind the reader of the formal definition of an equivalence relation.

**Definition 6.** Let $S$ be a set. Given $s$ and $t$ in $S$, we write $s \simeq t$ ($s$ is *equivalent to* $t$) when

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

We have already encountered one example of an equivalence relation. Namely, graph isomorphism is an equivalence relation on the set of graphs. The three demands in this definition are respectively called the reflexive, symmetric, and transitive properties of an equivalence relation. Together, these properties ensure that $S$ can be partitioned into disjunct subsets, where each subset consists of elements that are equivalent to each other, but not to any member of another subset in the partition.

Hence, if we find a meaningful equivalence relation between nodes, we can measure the size of equivalence classes and use that as a measure of anonymity. Let us begin by defining the following relation between nodes.

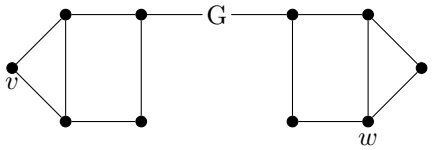
**Definition 7.** Let $G$ be a graph, $V$ its node set and $j$ a nonnegative integer or $\infty$. For a pair of nodes $v$ and $w$ in $V$ we write $v \simeq_j w$ when

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

If $v$ and $w$ are in $\simeq_j$ we shall write $v \simeq_j w$.

It is not hard to demonstrate (See Appendix A.1) that $\simeq_j$ is indeed an equivalence relation, as desired. The interpretation of this relation is that $v$ is equivalent to $w$ (in order $j$) when they have similar (isomorphic) neighbourhoods of order $j$ *and* when they play the same role in their respective neighbourhoods.

**Example 8.** Consider the following graph (where $G$ is an arbitrary graph).




(5)

The nodes $v$ and $w$ have isomorphic second order neighbourhoods, so $N(w, 2) \simeq N(v, 2)$. But there is no isomorphism $\phi : N(v, 2) \to N(w, 2)$ for which $\phi(v) = w$ (note for example that $v$ and $w$ have different degrees). So although $v$ and $w$ have isomorphic second order closed neighbourhoods, they do not satisfy the $\simeq_2$ relation.

We shall denote the subset of nodes equivalent to a particular node $v$ as $[v]_j$, specifically:

$$[v]_j = \{w \in V : v \simeq_j w\}. \tag{6}$$

We are now ready to formally define topological anonymity.

**Definition 9** (Topological anonymity)**.** Let $G$ be a graph, $v$ one of its nodes, and $j$ a nonnegative integer or $\infty$. The *$j$th order topological anonymity* $a(v,j)$ of $v$ is defined as the size of its equivalence class under the $\simeq_j$ relation

$$a(v,j) = \#[v]_j.$$

We now like to check whether $a(v,j)$ can be interpreted as the inverse of the disclosure probability $p(v|j)$ (see Equation 2) where $j$ quantifies the amount of information $I$ obtained by an adversary. In particular we would like to know whether $p(v|j+1) \geq p(v|j)$, as desired. It turns out that this property follows immediately from the following, more general Theorem.

**Theorem 10.** Let $G$ be a graph, $V$ its node set and $\delta = \mathrm{diam}(G)$ the graph diameter. For any $v \in V$ we have

$$\mathrm{Orbit}(v) = [v]_\delta \subseteq [v]_{\delta-1} \subseteq \cdots \subseteq [v]_1 \subseteq [v]_0 = V.$$

The proof is given in Appendix A.2. The main argument for proving the inclusions involves restricting an appropriate isomorphism $N(v,j+1) \mapsto N(w,j+1)$ to an appropriate isomorphism $N(v,j) \mapsto N(w,j)$. This Theorem will be important when we develop algorithms to compute equivalence classes. Here, we focus on the consequence that $a(v,j)$ is monotonically decreasing as a function of $j$.

**Corollary 11.**

$$\#\mathrm{Orbit}(v) = a(v,\delta) \leq a(v,\delta-1) \leq \cdots \leq a(v,1) \leq a(v,0) = n.$$

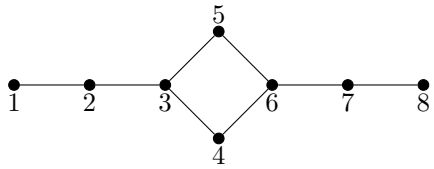
*Proof.* This follows immediately from Definition 9 and Theorem 10. □

Equation (2) can now be interpreted as

$$r(v,j) = p(v|j)p(j) = \frac{p(j)}{a(v,j)}, \tag{7}$$

where $p(j)$ is the probability that an adversary knows the structure of the neighbourhood of $v$ up to order $j$. We see that $p(v|0) = 1/n$. In other words, if an adversary knows nothing about a node except that it is part of the network, in which case $p(0) = 1$, then all nodes are equally likely. We see that as an adversary learns more about a wider neighbourhood of $v$, it's anonymity decreases, until it reaches the minimum defined by the number of elements in its orbit.

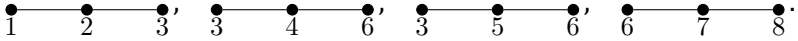
**Example 12** (Continued from Example 5)**.** Consider again the following graph $G$.




As an example, we determine the anonymity of node 2 as a function of $j$. At $j = 0$, we have the trivial case $[2]_0 = \{1, 2, 3, 4, 5, 6, 7, 8\}$ so $a(2, 0) = 8$. For $j = 1$, we need to find all occurrences of the pattern


(8)

such that there is an isomorphism mapping 2 to $v$. We find there are four such occurrences, namely:




(9)

So we have $[2]_1 = \{2, 4, 5, 7\}$ and thus $a(2, 1) = 4$. Similarly, we can compute $[2]_2 = \{2, 7\}$ and $a(2, 2) = 2$. Since $[2]_2 = \mathrm{Orbit}(2)$, Theorem 10 tells us that $a(2, j) = 2$ for $j = 2, 3, \ldots, 6$. The following table summarizes the value of $a(v, j)$ for all nodes $v$ and all relevant values of $j$.

| $a(v,j)$ | $j$ | | | | | | |
|---|---|---|---|---|---|---|---|
| $v$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 8 | 2 | 2 | 2 | 2 | 2 | 2 |
| 2 | 8 | 4 | 2 | 2 | 2 | 2 | 2 |
| 3 | 8 | 2 | 2 | 2 | 2 | 2 | 2 |
| 4 | 8 | 4 | 2 | 2 | 2 | 2 | 2 |
| 5 | 8 | 4 | 2 | 2 | 2 | 2 | 2 |
| 6 | 8 | 2 | 2 | 2 | 2 | 2 | 2 |
| 7 | 8 | 4 | 2 | 2 | 2 | 2 | 2 |
| 8 | 8 | 2 | 2 | 2 | 2 | 2 | 2 |

We see that nodes 1, 3,6, and 8 run a larger risk of being exposed then nodes 2, 4, 5 and 7, since knowing just their neighbourhood up to order 1, increases their disclosure probability to 1/2. For nodes 2, 4, 5, and 7, the disclosure risk is increased to 1/4 when an adversary obtains information about the structure of their 1st order neighbourhood.

Summarizing, we find that $a(v, j)$ can be interpreted as a measure of anonymity of a node $v$ that is induced by the neighbourhood of $v$ up to and including distance $j$. $a(v, j)$ decreases monotonically with increasing $j$ and has natural limits at $j = 0$ and $j = \mathrm{diam}(G)$. The inverse of $a(v, j)$ can be interpreted as the disclosure probability $p(v|j)$, where $j$ measures the amount of structural information obtained by an adversary in advance.

## 3.3 Anonymity for general neighbourhoods

The anonymity measure defined thus far yields disclosure probability for a target node when an adversary knows the complete neighbourhood of a target node up to and

including a certain distance. But what happens when an adversary doesn't know the complete neighbourhood? Here we generalize the concept of neighbourhood to any subgraph of the network that contains the target node. As before we start by defining a notion of equivalence.

**Definition 13.** Let $G$ be a graph and $A$ a subgraph such that $v$ is a node of $A$. We write $v \simeq_A w$ when

- there is a subgraph $A'$ of $G$ containing $w$ and $A' \simeq A$; and
- there is an isomorphism $\phi : V(A) \to V(A')$ such that $\phi(v) = \phi(w)$.

It can again be demonstrated that $\simeq_A$ is an equivalence relation. Hence, given any subgraph $A$ of $G$ containing $v$, we can define an equivalence class as follows:

$$[v]_A = \{w \in V : w \simeq_A v\}. \tag{10}$$

For a chosen target node $v$ and a chosen $A$, this separates $V$ into two disjunct classes: those nodes that are $A$-equivalent to $v$ and those that are not. Equivalently, we see that each subgraph $A$ generates $|V(A)|$ binary partitions of $V(G)$: one for each node in $A$.

This allows us to define a generalized anonymity measure as the cardinality of $[v]_A$:

$$a(v, A) = \#[v]_A. \tag{11}$$

Analogous to Theorem 10, we can find some limits for $[v]_A$ and hence for $a(v, A)$. The smallest subgraph containing $v$ occurs when $A = (\{v\}, \{\}) = N(v, 0)$. The largest subgraph containing $v$ occurs when $A = G = N(v, \mathrm{diam}(G))$. Furthermore, we have the following.

**Theorem 14.** Let $A$ and $B$ be subgraphs of $G$ containing a node $v$, such that $A$ is a subgraph of $B$. In this case we have $[v]_A \subseteq [v]_B$.

The proof is given in Appendix A.3. This is a generalization of Theorem 10, since we can always choose $A$ and $B$ to be neighbourhoods of $v$. We immediately obtain that if $A$ and $B$ are subgraphs of $G$ containing $v$ such that $A \subseteq B$, then $a(v, B) \leq a(v, A)$. This property can be used to interpolate between $a(v, j)$ and $a(v, j + 1)$ in the following way.

Consider a node $v$ and a surrounding neighbourhood $N(v, j)$. Every node in $N(v, j + 1)$ is connected to at least one node in $N(v, j)$. We generate a sequence of graphs $N(v, j), A, A', \dots, N(v, j + 1)$ by adding a single node and the edges that connect it to the nodes in the previous graph in each step. Each previous graph is a subgraph of the next, and hence we obtain a sequence of anonymity values
$a(v, j) \geq a(v, A) \geq a(v, A') \geq \cdots \geq a(v, j + 1)$.

In short, we see that the definition of anonymity developed in Section 3.2 can be extended to general surroundings of a node. The inclusion Theorem (Theorem 14) suggests that the anonymity values $a(v, j)$ can be interpreted as 'anonymity land marks'. This in the sense that if an adversary knows that a target node is part of a subgraph $A$, there is a smallest $N(v, j)$ containing $A$, and a largest $N(v, j')$ contained by $A$, and we have $a(v, j) \leq a(v, A) \leq a(v, j')$. The rest of this paper focuses on computing $a(v, j)$.

# 4 Algorithms

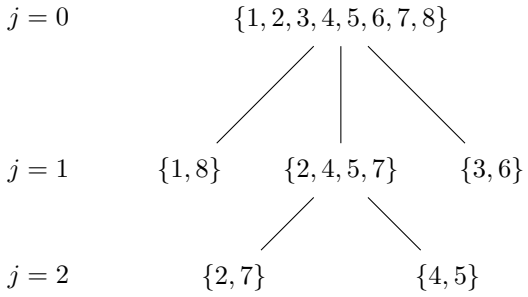
We now turn to the question on how to compute anonymity values for nodes of a graph as a function of $j$. For $j = 0$, the calculation is trivial: $a(v, 0) = n$ ($n$ the number of nodes) for all nodes $v$. For $j \geq \mathrm{diam}(G)$, we need to find the orbits of nodes in the graph. One approach is to first find all automorphisms, and then apply Equation 4 to generate the orbits. This problem is notoriously hard: generating all automorphisms could take exponential time, and it is currently unknown whether a polynomial method can be developed to find a smaller set of *generating elements* of the automorphism group. A generating set of elements is a subset of $\mathrm{Aut}(G)$ that recreates $\mathrm{Aut}(G)$ by multiplying the generating elements in different combinations and orders. It can be shown (e.g. Arvind (2007)) that finding a generating set for $\mathrm{Aut}(G)$ is equivalent to the graph isomorphism problem. Babai (2015) claims in a yet unconfirmed proof that it can be solved in quasipolynomial time. An recent overview of confirmed graph isomorphism approaches is given by McKay and Piperno (2014) and a recent discussion of algorithms for finding orbits is given by Mowshowitz and Mitsou (2009).

A natural approach to computing anonymity values for all nodes and all $j$ is to exploit that two nodes can only be equivalent with respect to $\simeq_{j+1}$, when they are equivalent with respect to $\simeq_j$ (Theorem 10). Algorithm 1 shows one way to organize the administration when splitting a $\simeq_j$ equivalence class of nodes into a set of $\simeq_{j+1}$ equivalence classes. The approach to compute all equivalence classes is then to start with the set $[v]_0$, and to apply Algorithm 1 recursively to each element of the output until no further split can be achieved. This is the approach taken to compute the results shown in Section 5.

**Algorithm 1:** Split an equivalence class of order $j$ into a set of $j+1$ equivalence classes.

**Input**: An equivalence class $[v]_j$.
**Result:** A partition $U$ of $[v]_j$, where each element of $U$ is a $j+1$ equivalence class.
$U \leftarrow \{\}$
**while** $[v]_j \neq \{\}$ **do**
    Choose a $v'$ from $[v]_j$
    **for** $[u]_{j+1} \in U$ **do**
        **if** $v' \simeq_{j+1} u$ **then**
            $[u]_{j+1} \leftarrow [u]_{j+1} \cup \{v'\}$ /* Add $v'$ to equiv class */
            $[v]_j \leftarrow [v]_j - \{v'\}$
            **break**
    **if** $v' \in [v]_j$ **then**
        $U \leftarrow U \cup \{\{v'\}\}$ /* Start new equiv class $[v']_{j+1}$ */
        $[v]_j \leftarrow [v]_j - \{v'\}$

**Example 15.** Consider again the graph $G$ of Examples 5 and 12. Starting at $j = 0$ we get $U = \{1, 2, \ldots, 8\}$. This is split into three equivalence classes using Algorithm 1. One of the resulting classes ($[2]_1$) is split again into two classes. This yields a tree structure as shown below (we suppress branches that do not yield any new splits).




In this simple example, we can stop the recursive application of Algorithm 1 because the orbits of all nodes are known. In practice this will not be the case. We can stop the recursion when either $j$ reaches the diameter of the largest component of $G$, or when a node ends up in a singleton set: i.e. when $[v]_j = \{v\}$.

# 5 An example

Algorithm 1 was implemented in Python, and results subsequently analyzed with R and `igraph` (R Core Team, 2020; Csardi and Nepusz, 2006). The main difficulty of this method is in line 6, where it is established whether two nodes are equivalent in order $j + 1$. We used the `NetworkX` module of Haghberg et al. (2019) to determine whether two neighbourhoods are isomorphic, and to compute explicit isomorphisms between neighbourhoods, when they are.

For reasons to be discussed, the current implementation takes a long time to run, even for networks of moderate size. For this reason we present a calculation on a scale-free network with 100 nodes, 2 edges per added node, and scale parameter $\alpha = 2$[4)]. We computed the equivalence classes up to and including $j = 1$.

The resulting equivalence classes are summarized in Figures 5.2. In the following table, we summarize how many nodes have a certain anonymity value. We also list into how many different classes each set of nodes with the same anonymity occur.

| $a(v, 1)$ | $p(v\|j)$ | #nodes | #classes |
|---:|---:|---:|---:|
| 1 | 1.000 | 10 | 10 |
| 2 | 0.500 | 4 | 2 |
| 4 | 0.250 | 4 | 1 |
| 5 | 0.200 | 5 | 1 |
| 7 | 0.143 | 7 | 1 |
| 15 | 0.067 | 15 | 1 |
| 55 | 0.018 | 55 | 1 |

[4)] A scale-free network is constructed by adding nodes and connecting them with $m$ (here: 2) links to existing nodes. The probability of connecting to an existing node is proportional to the number of connections it already has. The result is a network where the probability that a node has degree $k$ is proportional to $k^{-\alpha}$, in expectation. See e.g. Newman (2018, Chapter 14) for a textbook discussion.

Here, each row lists the computed anonymity, the disclosure risk, the number of nodes having this anonymity, and the number of different classes with this anonymity (given by $\#\text{nodes}/a(v,1)$). For example, there are 10 nodes that are unique for $j = 1$. Since each node is alone in its equivalence class, there are 10 such classes, and each such node is fully re-identified ($p(v|j) = 1$) when an adversary knows the structure of its first order neighbourhood.

Figure 5.1 shows the unique nodes and their respective first order neighbourhoods. Unique nodes include the nodes with highest degree, as expected, but there are a few nodes with neighbourhoods that do not seem special in any way. This may be an artefact of the small network size and low number of edges. Nevertheless, it is clear that these nodes run the highest risk of being re-identified based on knowledge of their surrounding network structure.

Figure 5.2 summarizes the nodes with higher anonymity. The largest anonymity class contains 55 nodes: more than half of the nodes in this network have a triangle as their first order neighbourhood. These nodes have the highest anonymity and thus the least change of being re-identified based on knowing their first-order neighbourhood: $p(v|j) = 1/55 \approx 0.018$.

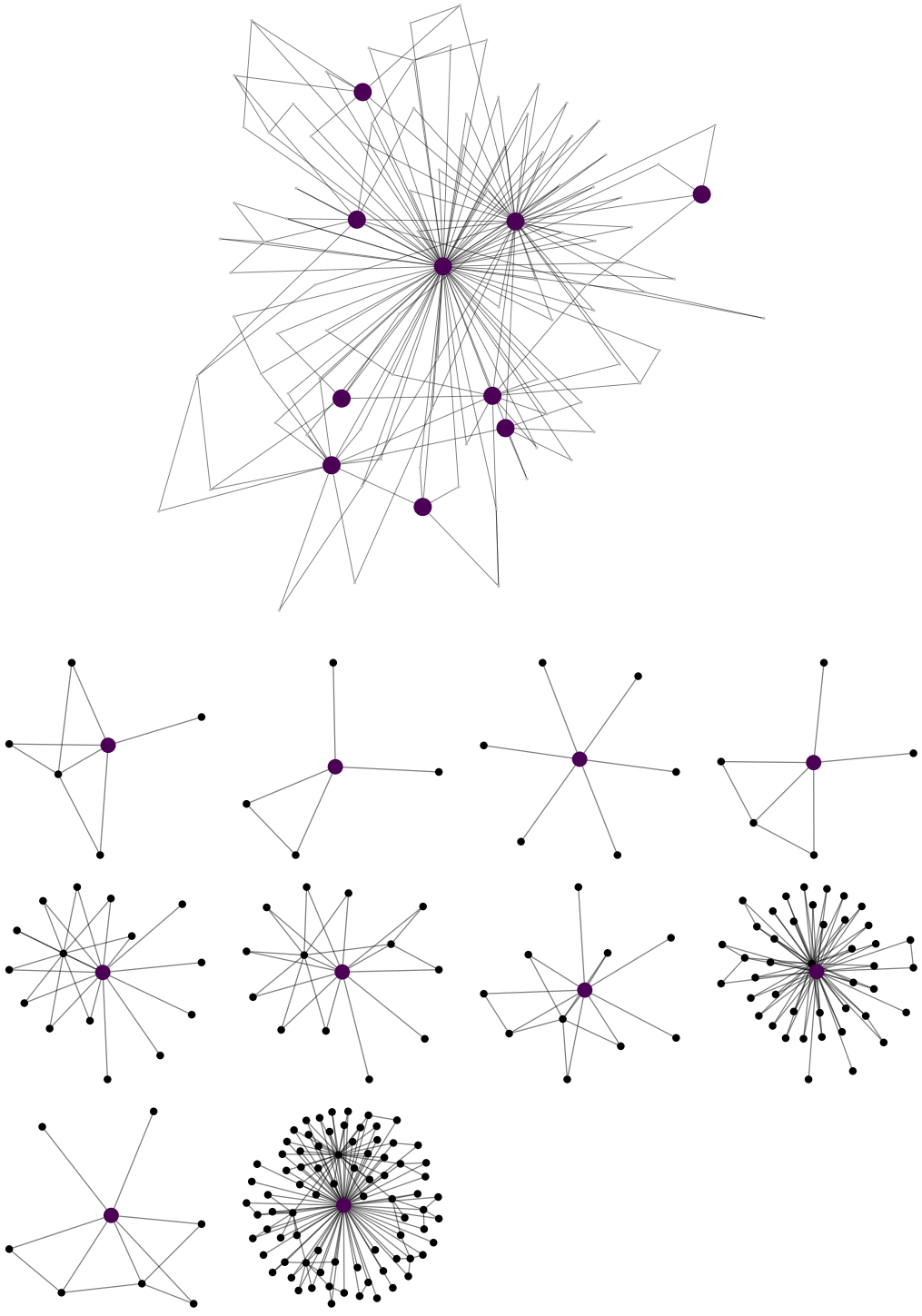


**Figure 5.1 Top: nodes that are unique in order $j = 1$. Bottom: the unique nodes in their 1st order neighbourhoods.**

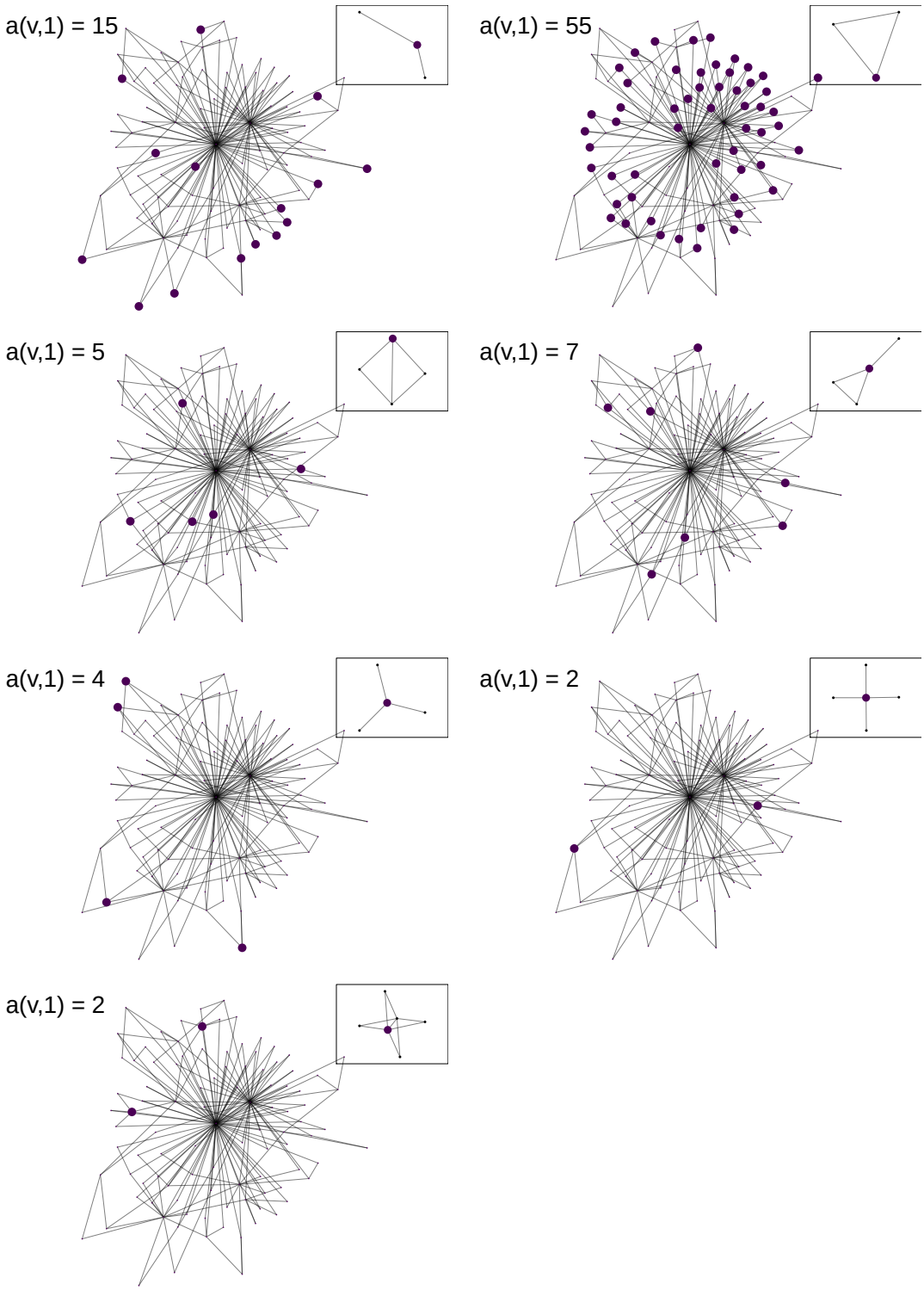


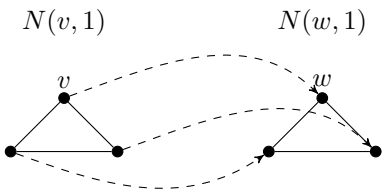

**Figure 5.2 Equivalent node sets in a scale-free network. Anonymity values are at the top left, the insets show the prototype neighbourhood and central node.**

# 6 Discussion and possible ways forward

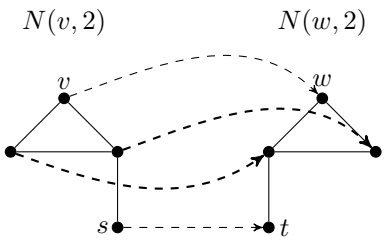
The current implementation of computing anonymity values is naive, in the sense that for each value of $j$, and for each pair of candidates $v$ and $w$, the neighbourhoods are derived, it is determined whether they are isomorphic, and if they are, it is determined whether there is an isomorphism mapping $v$ to $w$. In the current implementation this means that a lot of calculations are done and redone. The purpose of this section is to point out a few possible ways that would allow for quicker calculations in realistic networks.

One way forward is to keep an administration of isomorphisms, as we let the neighbourhoods surrounding two compared nodes grow. The inclusion Theorem 10, relies on the fact that if we have an isomorphism $N(v, j+1) \to N(w, j+1)$ that maps $v$ to $w$, we can by restriction find an isomorphism $N(v, j) \to N(w, j)$ that maps $v$ to $w$. We can wonder if we can also go the other way around: perhaps it is possible to start with an isomorphism $N(v, j) \to N(w, j)$ and *extend* it to $N(v, j+1) \to N(w, j+1)$. It turns out, we can, but there is a caveat.

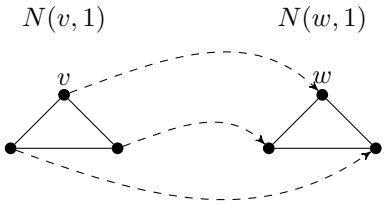
Consider the following situation, where we've found that $v$ and $w$ are equivalent to order $j = 1$. Below we depict an example, where the dashed arrow represent an isomorphism $\phi$ connecting the two neighbourhoods and where $\phi(v) = w$.



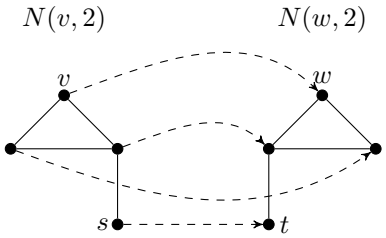

Now suppose we move on to $j = 2$. We find again that $N(v, 2) \simeq N(w, 2)$ and attempt to extend $\phi$ by adding the mapping between node $s$ and $t$ as shown below.




However, we now discover that part of what was originally an isomorphism, now maps between nodes of different degrees. These mappings are depicted here with thick arrows. The solution is to go back, and choose another isomorphism between $N(v, 1)$ and $N(w, 1)$ that maps $v$ to $w$.




And indeed, this one can be extended successfully.




In this case, the 'mistake' in choosing an isomorphism to extend was discovered at the first extension, moving from $j$ to $j+1$. However in the general case we can not exclude the possibility that such a wrong turn is discovered only at a later step. This suggests a strategy where we explore possible paths of extending (sequences of ) isomorphisms that map $v$ to $w$.

A second, and more straightforward way to gain performance is to use the fact that Algorithm 1 can be run in parallel over equivalence classes. This follows from the fact that nodes can only be equivalent in order $j+1$ when they are equivalent in order $j$. A third way forward is to use a compiled language in stead of a scripting language such as Python.

It is worth noting that these algorithms are in close relation to the graph isomorphism problem. In fact, Algorithm 1 described here turns out to be similar to what McKay and Piperno (2014) in a review of graph isomorphism approaches call a 'refinement function'. The complexity of determining whether two graphs are isomorphic is a famous open problem, although a recent, yet unconfirmed claim of Babai (2015) states that it can be solved in quasipolynomial time. Another area that is interesting in this respect is the literature on graphlet analyses. A graphlet is a small connected graph, and in graphlet analyses a node is characterized by the graphlets in which it appears. See Sarajlić et al. (2016); Rahman (2016), Hočevar et al. (2016) for some recent references. Graphlet analyses may also offer ways to approximate anonymity values rather than computing them exactly.

# 7 Summary and conclusion

Network data offers interesting ways for official statisticians and scientists to investigate society and economy. Making anonymized network microdata available for research or

publication presents new challenges for Statistical Disclosure Control, since network structure may facilitate re-identification of nodes. In order to assess the risk of re-identification, a measure of non-uniqueness, or anonymity of nodes with regard to surrounding network structure is necessary.

In this paper we have defined a measure of anonymity that is based on counting the number of nodes that have the same surrounding neighbourhood, while playing the same role in that neighbourhood. It was demonstrated that this measure has a few pleasant properties: the notion of anonymity increases as more of a node's surrounding network structure is known, and there are natural limits, in cases where the complete graph is known or when nothing is known of a node's surroundings.

Moving forward, there are a number of open issues. The first practical issue is that we currently have no algorithm allowing for fast calculation of anonymity values. In this paper we have pointed out a few possible ways to make progress, including improving the current approach and perhaps searching for ways to approximate anonymity.

A second issue regards the statistical properties of anonymity, in different network models. For example, it would be interesting to know which levels of anonymity one may hope to achieve, given that a network has the properties of a scale-free, or Erdős-Rényi network.

Third, the question on how to improve anonymity in a network is as of yet open. At the moment we have no methodology to perturb or suppress parts of a network data while keeping important statistical properties intact.

Fourth, from the point of view of disclosure control there is the question of anonymity in the case where the data does not cover the population. That is, the data may cover only a (sampled) part of the population. How to translate such situations to the case of network data is not discussed in this work —we have silently assumed that the available network data covers the population.

Finally, there may be other attack scenarios based on structural properties, such as a node's betweennes or centraelity. These have not been treated here.

# References


Arvind, V. (2007). Algebra and computation. Lecture notes (transcribed by Ramprasad Saptharishi), `http://www.cmi.ac.in/~ramprasad/lecturenotes/algcomp/tillnow.pdf` (last viewed 15-06-2020).

Babai, L. (2015). Graph isomorphism in quasipolynomial time. *CoRR abs/1512.03547*.

Csardi, G. and T. Nepusz (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, 1695.

de Jong, R., M. van der Loo, and F. Takes (2021a). Measuring anonymity in complex networks. In *Networks2021: A joint Sunbelt and NetSci conference*.

de Jong, R., M. van der Loo, and F. Takes (2021b). Measuring anonymity in complex networks. In *IC2S2 2021: 7th international conference on computational social science*.

de Jong, R. G. (2021). Measuring structural anonymity in complex networks. Msc thesis, Leiden Institute of Advanced Computer Science, Leiden University, Niels Bohrweg 1, 2333CA Leiden, The Netherlands.

Diestel, R. (2000). *Graph theory*. Graduate texts in mathematics. Springer-Verlag New York, Incorporated.

Haghberg, A., D. Schult, and P. Swart (2019). *NetworkX Reference*. Release 2.4.

Hay, M., G. Miklau, D. Jensen, P. Weis, and S. Srivastava (2007). Anonymizing social networks. Computer Science Department Faculty Publication Series 180, University of Massachusetts Amherst.

Hočevar, T., J. Demšar, et al. (2016). Computation of graphlet orbits for nodes and edges in sparse graphs. *Journ. Stat. Soft 71*.

Hundepool, A., J. Domingo-Ferrer, L. Franconi, S. Giessing, E. Schulte Nordholt, K. Spicer, and P.-P. De Wolf (2012). *Statistical disclosure control*. John Wiley & Sons.

Ji, S., P. Mittal, and R. Beyah (2016). Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey. *IEEE Communications Surveys & Tutorials 19*(2), 1305–1326.

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

Mowshowitz, A. and V. Mitsou (2009). Entropy, orbits and spectra of graphs. *Analysis of Complex Networks: From Biology to Linguistics, Wiley-VCH*.

Newman, M. (2018). *Networks*. Oxford university press.

Newman, M. E. (2011). Complex systems: A survey. *Am. J. of Physics 79*, 800–810.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rahman, M. (2016). *Graphlet based network analysis*. Ph. D. thesis, Purdue University.

Samarati, P. (2001). Protecting respondents identities in microdata release. *IEEE transactions on Knowledge and Data Engineering 13*(6), 1010–1027.

Samarati, P. and L. Sweeney (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression.

Sarajlić, A., N. Malod-Dognin, Ö. N. Yaveroğlu, and N. Pržulj (2016). Graphlet-based characterization of directed networks. *Scientific reports 6*, 35098.

van der Laan, J. and E. de Jonge (2017). Producing official statistics from network data. In *International Conference on Network Theory and Their Applications*, pp. 288.

van der Laan, J. and E. de Jonge (2019). Measuring local assortativity in the presence of missing values. In *International Conference on Complex Networks and Their Applications*, pp. 280–290. Springer.

Willenborg, L. and T. De Waal (2001). *Elements of statistical disclosure control*, Volume 155. Springer Science & Business Media.

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.

# Appendix

# A Proofs

## A.1 Equivalence relation for nodes

*Proof.* We need to show that $\simeq_j$ is reflexive, symmetric, and transitive. Let $v, w, u$ be nodes of $V$.

Reflexivity: we have $N(v,j) \simeq N(v,j)$ and we can choose the identity function as the isomorphism mapping $v$ to $v$.

Symmetry: if $v \simeq_j w$ then $N(v,j) \simeq N(w,j)$ so $N(w,j) \simeq N(v,j)$ by definition of graph isomorphism. If $v \simeq_j w$ then there is a bijection $\phi$ sending $v$ to $w$ and we can choose the inverse of $\phi$ as the isomorphism mapping $w$ to $v$.

Transitivity: suppose $v \simeq_j w \wedge w \simeq_j u$. By transitivity of graph isomorphism we have $N(v,j) \simeq N(u,j)$. Now, let $\phi : N(v,j) \to N(w,j)$ be the isomorphism such that $\phi(v) = w$ and let $\theta : N(w,j) \to N(u,j)$ be the isomorphism such that $\theta(w) = u$. The composit $\theta\phi$ is an isomorphism $N(v,j) \to N(u,j)$ such that $(\theta\phi)(v) = u$. □

## A.2 Proof of Theorem 10

The hardest part is to prove the inclusion property. The intuition behind the proof is to show that two nodes can only be equivalent in order $j$ when they are equivalent in order $j-1$.

*Proof.* For the equality $\mathrm{Orbit}(v) = [v]_\delta$, observe that if $v \simeq_\delta w$ then $N(v,\delta) = N(w,\delta) = G$ so the isomorphism mapping $v$ to $w$ is an automorphism of $G$.

To demonstrate the inclusion properties, suppose that $v \simeq_j w, j > 0$. This means that there is an isomorphism $\phi : N(v,j) \mapsto N(w,j)$ such that $\phi(v) = w$. Now define $\phi' : N(v,j-1) \mapsto N(w,j-1)$ as the restriction of $\phi$ to $N(v,j-1)$. This is again an isomorphism with $\phi'(v) = w$, and hence $v \simeq_{j-1} w$. Thus, if $w \in [v]_j$ it must also be in $[v]_{j-1}$, as desired.

For the equality, $[v]_0 = V$, choose a particular $v \in V$. We have $N(v,0) = (\{v\}, \{\})$. Thus for each $w \in V$ there is a unique isomorphism $\phi^* : N(v,0) \mapsto N(w,0)$ such that $\phi^*(v) = w$. Hence, $v \simeq_0 w$ for all $w \in V$. □

## A.3 Proof of Theorem 14

The proof is very similar to the proof of the inclusion property of Theorem 10.

*Proof.* Suppose that $v \simeq_B w$, so that $w \in [v]_B$. This means that there is a subgraph $B'$ in $G$ containing $w$, such that $B \simeq B'$ and there is an isomorphism $\phi : V(B) \to V(B')$ such that $\phi(v) = w$. The restriction of $\phi$ to $V(A)$ (with $A \subseteq B$) is also an isomorphism. It maps nodes of $A$ to a subgraph $A'$ of $B'$. Since $v$ is an element of $V(A)$, we have again that $v$ maps to $w$. Hence: $v \simeq_A w$ and $[v]_B \subseteq [v]_A$ as desired. □

## Colophon

*Publisher*
Statistics Netherlands
Henri Faasdreef 312, 2492 JP The Hague
www.cbs.nl

*Prepress*
Statistics Netherlands, Grafimedia

*Design*
Edenspiekermann

*Information*
Telephone +31 88 570 70 70, fax +31 70 337 59 94
Via contact form: www.cbs.nl/information