Knowledge and Information Systems (2022) 64:1077–1100
https://doi.org/10.1007/s10115-022-01662-z

REGULAR PAPER

Preventing active re-identification attacks on social graphs via sybil subgraph obfuscation

Sjouke Mauw[1] · Yunior Ramírez-Cruz[1] · Rolando Trujillo-Rasua[2]

Received: 10 July 2020 / Revised: 17 January 2022 / Accepted: 21 January 2022 /
Published online: 27 February 2022

Abstract

Active re-identification attacks constitute a serious threat to privacy-preserving social graph publication, because of the ability of active adversaries to leverage fake accounts, a.k.a. *sybil nodes*, to enforce structural patterns that can be used to re-identify their victims on anonymised graphs. Several formal privacy properties have been enunciated with the purpose of characterising the resistance of a graph against active attacks. However, anonymisation methods devised on the basis of these properties have so far been able to address only restricted special cases, where the adversaries are assumed to leverage a very small number of sybil nodes. In this paper, we present a new probabilistic interpretation of active re-identification attacks on social graphs. Unlike the aforementioned privacy properties, which model the protection from active adversaries as the task of making victim nodes indistinguishable in terms of their fingerprints with respect to all potential attackers, our new formulation introduces a more complete view, where the attack is countered by jointly preventing the attacker from retrieving the set of sybil nodes, and from using these sybil nodes for re-identifying the victims. Under the new formulation, we show that k-symmetry, a privacy property introduced in the context of passive attacks, provides a sufficient condition for the protection against active re-identification attacks leveraging an arbitrary number of sybil nodes. Moreover, we show that the algorithm K- MATCH, originally devised for efficiently enforcing the related notion of k-automorphism, also guarantees k-symmetry. Empirical results on real-life and synthetic graphs demonstrate that our formulation allows, for the first time, to publish anonymised social graphs (with formal privacy guarantees) that effectively resist the strongest active re-identification attack reported in the literature, even when it leverages a large number of sybil nodes.

Keywords Private social graph publication · Anonymisation · Active adversaries

✉ Yunior Ramírez-Cruz
yramirezc@gmail.com

Sjouke Mauw
sjouke.mauw@uni.lu

Rolando Trujillo-Rasua
rolando.trujillo@deakin.edu.au

[1] DCS, SnT, University of Luxembourg, Esch-sur-Alzette, Luxembourg

[2] School of Information Technology, Deakin University, Burwood, VIC, Australia

1 Introduction

The last decade has witnessed a formidable explosion in the use of social networking sites. Although the discipline of social network analysis has existed already for quite some time, today's scientists potentially have access as never before to massive amounts of social network data. Social graphs are a particular example of this type of data, in which vertices typically represent users (e.g. Facebook or Twitter users, e-mail addresses) and edges represent relations between these users (e.g. becoming "friends", following someone, exchanging e-mails). The analysis of social graphs can help scientists and other actors to discover important societal trends, study consumption habits, understand the spread of news or diseases, etc. For these goals to be achievable, it is necessary that the holders of this information, e.g. online social networks, messaging services, among others, release samples of their social graphs. However, ethical considerations, increased public awareness and reinforced legislation[1] place an increasingly strong emphasis on the need to protect individuals' privacy via anonymisation.

Social graphs have proven themselves a challenging data type to anonymise. Even a simple undirected graph, with arbitrary node labels and no attributes on vertices or edges, is susceptible of leaking private information, due to the existence of unique structural patterns that characterise some individuals, e.g. the number of friends or the relations in the immediate vicinity [35]. Many privacy attacks that solely rely on the underlying graph topology of the social graph exist [1], and they are still effective [32], despite advances on social graph anonymisation. A particularly effective privacy attack is the so-called *active attack*, which uses a strategy consisting in inserting fake accounts, commonly referred to as *sybils*, into the real network. Once inserted, these fake users interact with legitimate users and among themselves, and create structures that allow the adversary to retrieve the sybil nodes from a sanitised social graph and use the connection patterns between sybils and legitimate nodes to re-identify the original users and infer sensitive information about them, such as the existence of relations.

The publication of social graphs that effectively resist active attacks was initially addressed by Trujillo-Rasua and Yero [46]. They introduced the notion of (k, ℓ)-anonymity, the first privacy property to explicitly model the protection of published graphs against active adversaries. A graph satisfying (k, ℓ)-anonymity ensures that an adversary leveraging up to ℓ sybil nodes and knowing the pairwise distances of all victims to all sybil nodes, is still unable to distinguish each victim from at least $k - 1$ other vertices in the graph. This privacy property served as the basis for defining several anonymisation methods for a particular case, namely the one where either $k > 1$ or $\ell > 1$ [30,33]. In other words, non-trivial anonymity ($k > 1$) was only guaranteed against an adversary leveraging exactly one sybil node. Later, the introduction of the notion of (k, ℓ)-adjacency anonymity [31] allowed to arbitrarily increase the values of k for which a formal privacy guarantee can be provided, but the proposed methods remained unable to address scenarios where the adversary can leverage more than two sybil nodes. In consequence, until now no anonymisation method with theoretically sound privacy guarantees against active attackers leveraging three or more sybil nodes has been made available to data publishers. This article solves such problem.

Our solution consists of identifying and formalising a more precise privacy model for active attacks, in terms of the capabilities the adversary is supposed to have, than those existing in the literature. We remove the assumption that the adversary is always capable

[1] For example, the European GDPR, which can be consulted at https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en.

of identifying the set of sybil nodes in the published graph, which appears in all privacy properties for active attacks [31,46] that we are aware of. In this new model, instead, the analyst needs to calculate the actual probability of success of the attacker on re-identifying the sybil nodes and combines it with the attacker's probability of re-identifying the victims.

By studying active attacks without the assumption that the attacker first needs to re-identify sybil nodes, we reached two main results: one of practical interest and another one theoretical. Of practical interest is our proof that the algorithm K- MATCH [54], originally devised for efficiently enforcing the notion of k-automorphism, makes it impossible for an active attacker to re-identify a victim with probability higher than $1/k$, regardless of the adversary strength. Hence we show K- MATCH to be the first anonymisation method that protects against active attackers of arbitrary strength. Second, we prove our privacy model to be a proper extension of previous models [31,46,50], in the sense that it describes all graphs that have been previously considered private, and describes others that are not captured by existing models. This allowed us to establish the first connection between privacy models for passive attacks, such as k-symmetry [50], and privacy models for active attacks. For example, we prove that k-symmetry and (k, ℓ)-anonymity are mutually exclusive, yet they are both proper instances of our privacy model. In other words, both models are sound, as far as resistance to active attack is concerned, but not complete. Whether there exists a k-anonymity model that captures all graphs resistant to active attacks, i.e. that is complete, is an open question.

Summary of contributions:

- We show that no privacy property in the literature characterises all anonymous graphs with respect to active attacks.
- We introduce a general definition of resistance to active attacks that can be used to analyse the actual resistance of a graph.
- We use the introduced privacy model to prove that k-symmetry, the strongest notion of anonymity against passive attacks, also protects against active attacks.
- Of independent interest is our proof that k-automorphism does not protect against active attacks. This is a surprising result, considering that k-automorphism and k-symmetry have traditionally been deemed as conceptually equivalent.
- We prove that the algorithm K- MATCH, devised to ensure a sufficient condition for k-automorphism, also guarantees k-symmetry.
- We provide empirical evidence on the effectiveness of K- MATCH as an anonymisation strategy against the strongest active attack reported in the literature, namely the robust active attack presented in [32], even when it leverages a large number of sybil nodes.

1.1 Structure of the paper

We discuss related work in Sect. 2 and describe our new probabilistic interpretation of the adversarial model for active re-identification attacks in Sect. 3. Then, we discuss the applicability of k-symmetry for modelling protection against active attackers in Sect. 4 and show in Sect. 5 that the algorithm K- MATCH efficiently provides a sufficient condition for k-symmetry. Finally, we empirically demonstrate the effectiveness of K- MATCH against the robust active attack from [32] in Sect. 6 and give our conclusions in Sect. 7.

2 Related work

In this paper, we focus on a particular family of properties for privacy-preserving publication of social graphs: those based on the notion of k-anonymity [43,45]. These privacy properties depend on assumptions about the type of knowledge that a malicious agent, the *adversary*, possesses. According to this criterion, adversaries can be divided into two types. On the one hand, *passive* adversaries rely on information that can be collected from public sources, such as public profiles in online social networks, where a majority of users keep unmodified default privacy settings that pose no access restrictions on friend lists and other types of information. A passive adversary attempts to re-identify users in a published social graph by matching this information to the released data. On the other hand, *active* adversaries not only use publicly available information, but also attempt to interact with the real social network before the data is published, with the purpose of forcing the occurrence of unique structural patterns that can be retrieved after publication and used for learning sensitive information.

2.1 *k*-anonymity models against passive attacks

k-anonymity is based on a notion of indistinguishability between users in a dataset, which is used to create equivalence classes of users that are pairwise indistinguishable to the eyes of an attacker. Formally, given a symmetric, reflexive and transitive indistinguishability relation $\sim$ on the users of a graph G, G satisfies k-anonymity with respect to $\sim$ if and only if the equivalence class with respect to $\sim$ of each user in G has cardinality at least k.

Several graph-oriented notions of indistinguishably appear in the literature. For example, Liu and Terzi [25] consider two users indistinguishable if they have the same degree. Their model is known as *k-degree anonymity* and gives protection against attackers capable of accurately estimating the number of connections of a user. The notion of k-degree anonymity has been widely studied, and numerous anonymisation methods based on it have been proposed, e.g. [5,6,12,26,27,39,41,48]. Zhou and Pei [53] assume a stronger attacker able to determine not only the connections of a user u, but also whether u's friends (i.e. those users that u is connected to) are connected. This means that the adversary is assumed to know the induced subgraphs created by the users and their neighbours. It is simple to see that Zhou and Pei's model, known as *k-neighbourhood anonymity*, is stronger than k-degree anonymity.

Another privacy notion that relies on the neighbourhood of a user is (k, ℓ)-anonymity [16], introduced by Feder, Nabar and Terzi and later generalised by Stokes and Torra [44]. In (k, ℓ)-anonymity, ℓ represents the number of neighbours two vertices ought to have to be considered indistinguishable. This indistinguishability relation is not transitive, though, making (k, ℓ)-anonymity hard to compare with other privacy properties based on neighbourhood, such as k-degree anonymity and k-neighbourhood anonymity.

The notion of *k-automorphism* [54] was introduced with the goal of modelling the knowledge of any passive adversary. Two users u and v in a graph G are said to be automorphically equivalent, or indistinguishable, if $\varphi(u) = v$ for some automorphism φ in G. The notion of k-automorphism ensures that every vertex in the graph is automorphically equivalent to $k - 1$ other vertices. Although k-automorphism itself does not in general imply all other privacy properties (as we will show in Appendix A), the method proposed in [54] for enforcing the (stronger) *k different matches principle* does achieve this goal. Similar formulations of indistinguishability in terms of graph automorphisms were presented independently in the work on k-symmetry [50] and k-isomorphism [11]. While k-symmetry and k-automorphism have traditionally been viewed as equivalent, k-symmetry is actually stronger, and it does

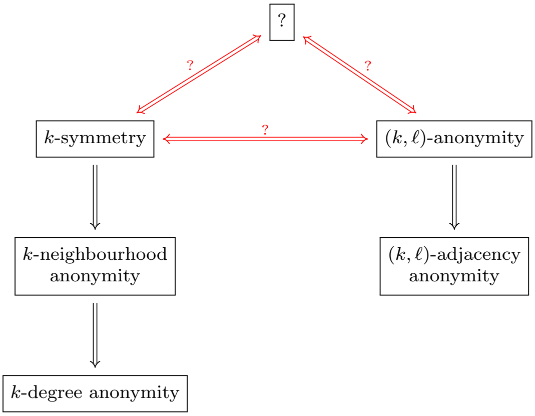


Fig. 1 A hierarchy of privacy properties. An arrow has the standard logical interpretation, i.e. $P \implies P'$ means that a graph satisfying P also satisfies P'. Left side: models for passive attacks. Right side: models for active attacks. Interrogation marks indicate connections that have not been established yet

imply all other privacy properties for passive attacks. In this paper, we additionally show that, in the context of active attacks, k-symmetry always guarantees a $1/k$ upper bound on the re-identification probability for each vertex, which k-automorphism does not.

A natural trade-off between the strength of the privacy notions and the amount of structural disruption caused by the anonymisation methods based on them has been empirically demonstrated in [54]. The three privacy models described above form a hierarchy, which is displayed in the left branch of Fig. 1. Privacy models tailored to active attacks also form a hierarchy, displayed in the right branch of Fig. 1, which we describe next. Interrogation marks in Fig. 1 indicate that connections between properties tailored for passive attacks and those tailored for active attacks have not been established yet, neither directly nor via some additional property.

2.2 *k*-anonymity models against active attacks

Backstrom et al. were the first to show the impact of active attacks in social networks back in 2007 [2]. Their attack has been optimised a number of times, see [32,37,38], and two privacy models particularly tailored to measure the resistance of social graphs to this type of attack have been recently proposed [31,46]. The first of those models is (k, ℓ)-*anonymity*, introduced in 2016 by Trujillo-Rasua and Yero [46]. They consider adversaries capable of re-identifying their own sets of sybil nodes in the anonymised graph. Adversaries are also assumed to know or able to estimate the distances of the victims to the set of sybil nodes. This last assumption was weakened later in [31] by restricting the adversary's knowledge to distances between victims and sybil nodes of length one. That is, the adversary only knows whether the victim is connected to a sybil node. That restriction led to a weaker version of (k, ℓ)-anonymity called (k, ℓ)-*adjacency anonymity*, as displayed in Fig. 1.

It is worth pointing out the clash in terminology with the use of (k, ℓ)-anonymity in [16] and [46]. Because this article focuses on active attacks, from now on whenever we write (k, ℓ)-anonymity we are referring to the privacy model that captures the resistance of a graph to active attacks, i.e. to that introduced in [46].

There exist three anonymisation algorithms [30,31,33] that aim to create graphs satisfying (k, ℓ)-(adjacency) anonymity. Their approach consists in determining a candidate set of sybil vertices in the original graph that breaks the desired anonymity property, and forcing via graph transformation that every vertex has a common pattern of connections with the sybil vertices shared by at least $k-1$ other vertices. A common shortcoming of these methods is that they only provide formal guarantees against attackers leveraging a very small number of sybil nodes (no more than two). This limitation seems to be an inherent shortcoming of the entire family of properties of which (k, ℓ)-anonymity and (k, ℓ)-adjacency anonymity are members. Indeed, for large values of ℓ, which are required in order to account for reasonably capable adversaries, anonymisation methods based on this type of property face the problem that any change introduced in the original graph to render one vertex indistinguishable from others, in terms of its distances to a vertex subset, is likely to render this vertex unique in terms of its distances to other vertex subsets.

2.3 Other privacy models

For the sake of completeness, we finish this brief literature review by surveying probabilistic privacy models. A popular example is differential privacy (DP) [13], a semantic privacy notion which, instead of anonymising the dataset, focuses on the methods accessing the sensitive data and provides a quantifiable privacy guarantee against an adversary who knows all but one entry in the dataset. In the context of graph data, the notion of two datasets differing by exactly one entry can have multiple interpretations, the two most common being *edge-differential privacy* and *vertex-differential privacy*. While a number of queries, e.g. degree sequences [18,22] and subgraph counts [21,52], have been addressed under (edge-)differential privacy, the use of this notion for numerous very basic queries, e.g. graph diameter, remains a challenge. Recently, differentially private methods leveraging the randomised response strategy for publishing a graph's adjacency matrix were proposed in [42]. While these methods do not necessarily view vertex ids as sensitive, data holders whose goal in preventing re-identification attacks is to prevent the adversary from learning the existence of relations may view this approach as an alternative to k-anonymity-based methods. Another DP-based alternative to k-anonymity-based methods consists in learning the parameters of a graph generative model under differential privacy and then using this model to publish synthetic graphs that resemble the original one in some structural properties [10,20,34,40,49,51].

Random perturbation for graph privacy has been used prior the introduction of differential privacy [7]. For example, within the context of passive attacks, Bonchi et al. [4] introduced a method that randomly removes and adds edges to the original graph. The anonymity level offered by their approach is evaluated against an information-theoretic measure that considers the uncertainty added to the original graph. We observe that randomisation techniques have not been successfully adapted to counteract active attacks. While intuition suggests that the task of re-identification becomes harder for the adversary as the amount of random noise added to a graph grows, it has been shown in [32] that active attacks can be made robust against reasonably large amounts of random perturbation.

Other probabilistic privacy models rely on the notion of *adversary's prior belief*, defined as a probability distribution on sensitive values. For example, t-closeness [24] measures attribute protection in terms of the distance between the distribution of sensitive values in the anonymised dataset with respect to the distribution of sensitive attribute values in the original table. Such definition of prior belief is different to other works, such as (ρ_1, ρ_2)-privacy [15] and ϵ-privacy [28], where prior belief represents the adversary's knowledge in the absence of knowledge about the dataset. In either case, estimating the prior belief of the adversary is challenging, as discussed in [13].

2.4 Concluding remarks

As illustrated in Fig. 1, the development of k-anonymity models against passive and active attacks has been traditionally split and had no apparent intersection. This article provides, to the best of our knowledge, the first connection between the two developments. This is achieved by introducing a probabilistic model for active attacks that characterises all graphs that resists active attacks, of which k-symmetry and (k, ℓ)-anonymity are proven to be sufficient, yet not necessary, conditions.

3 Probabilistic adversarial model

Our adversarial model is a generalisation of the model introduced in [32], which captures the capabilities of an active attacker and allows one to analyse the resistance of anonymisation methods to active attacks. Such analysis is expressed as a three-step game between the attacker and the defender. In the first step, the attacker is allowed to interact with the network, insert sybil accounts and establish links with other users (called the victims). The defender uses the second step to anonymise and perturb the network, which was previously manipulated by the attacker. Lastly, the attacker receives the anonymised network and makes a guess on the pseudonyms used to anonymise the victims. Each of these steps is formalised in what follows.

3.1 Attacker subgraph creation

The attacker–defender game starts with a graph $G = (V, E)$ representing a snapshot of a social network, as in Fig. 2a. The attacker knows a subset of the users, called the victims and denoted I, but not the connections between them. The attacker is allowed to insert a set of sybil nodes S into G and establish connections with their victims.

This step of the attack transforms the original graph $G = (V, E)$ into a graph $G^+ = (V', E')$ satisfying the following two properties: i) $V' = V \cup S$ and ii) $E' \setminus E \subseteq (S \times S) \cup (S \times I) \cup (I \times S)$. The second condition says that relations established by the adversary are constrained to the set of sybil and victim nodes. We call the resulting graph G^+ the *sybil-extended* graph of G. An example of a sybil-extended graph is depicted in Fig. 2b.

The attacker does not know the entire graph G^+, unless the original graph was empty. The adversary knows, however, the subgraph formed by the set of sybil nodes S, their connections to the victims, and the victim set I. This notion of adversary knowledge is formalised next.

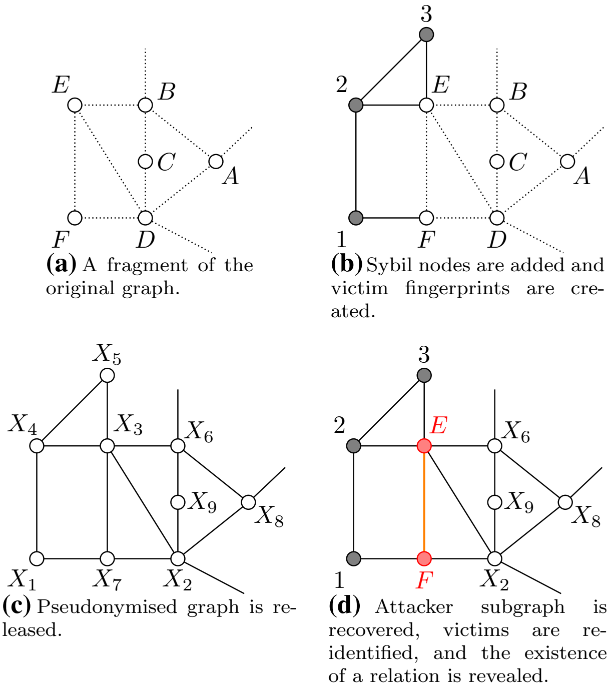
Definition 1 (*Adversary knowledge*) Let $G = (V, E)$ be an original graph and $G^+ = (V \cup S, E')$ the sybil-extended graph created by an adversary that targets a set of victims $I \subseteq V$. The *adversary knowledge* is defined as the subgraph $G_{S,I}$ of G defined by

$$G_{S,I} = (S \cup I, \{(u, v) \in E \mid \{u, v\} \subseteq S \cup I \wedge \{u, v\} \nsubseteq I\})$$

Note that connections between victims are not part of the adversary knowledge.

3.2 Pseudonymisation and perturbation

When the defender decides to publish the graph G^+, she pseudonymises it by replacing the real user identities with pseudonyms. That is to say, the defender obtains G^+ and constructs



(a) A fragment of the original graph.

(b) Sybil nodes are added and victim fingerprints are created.

(c) Pseudonymised graph is released.

(d) Attacker subgraph is recovered, victims are re-identified, and the existence of a relation is revealed.

Fig. 2 An active re-identification attack viewed as an attacker–defender game

an isomorphism φ from G^+ to φG^+. An *isomorphism* between two graphs $G = (V, E)$ and $G' = (V', E')$ is a bijective function $\varphi \colon V \to V'$, such that $\forall v_1, v_2 \in V \colon (v_1, v_2) \in E \iff (\varphi(v_1), \varphi(v_2)) \in E'$. Two graphs are *isomorphic*, denoted by $G \simeq_\varphi G'$, or briefly $G \simeq G'$, if there exists an isomorphism φ between them. Given a subset of vertices $S \subseteq V$, we will often use φS to denote the set $\{\varphi(v) | v \in S\}$. In Fig. 2c, we illustrate a pseudonymisation of the graph in Fig. 2b.

We call φG^+ the *pseudonymised* graph. Pseudonymisation serves the purpose of removing personally identifiable information from the graph. Because pseudonymisation is insufficient to protect a graph against re-identification, the defender is also allowed to perturb the graph. This is captured by a non-deterministic procedure t that maps graphs to graphs. The procedure t modifies φG^+, resulting in the *transformed* graph $t(\varphi G^+)$. We assume that $t(\varphi G^+)$ is ultimately made available to the public, hence it is known to the adversary.

3.3 Re-identification

The last step of the attacker–defender game is where the attacker analyses the published graph $t(\varphi G^+)$ to re-identify her own sybil accounts and the victims (see Fig. 2d). This allows her to acquire new information, which was supposed to remain private, such as the fact that E and F are friends.

We define the output of the adversary re-identification attempt as a mapping ρ from the set of vertices $S \cup I$ to the set of vertices in $t(\varphi G^+)$. This represents the adversary's belief on the pseudonyms used to anonymise the attacker and victim vertices in $t(\varphi G^+)$. To account for uncertainty on the adversary's belief, we consider that the adversary assigns a probability value $p(\rho)$ to each mapping, denoting the probability that the adversary chooses ρ as the output of the re-identification attack. Let $\Phi_{S,I}$ be the universe of mappings from the set of vertices in $S \cup I$ to the set $t(\varphi G^+)$. The law of total probability allows us to quantify the adversary's probability of success in re-identifying one victim as follows.

Proposition 1 *Let $G = (V, E)$ be an original graph, $G^+ = (V \cup S, E')$ the sybil-extended graph created by an adversary that targets a set of victims $I \subseteq V$, and $t(\varphi G^+)$ the anonymised version of G^+ created by the defender. Then, the probability $A^{S,I}_{t(\varphi G^+)}(u)$ that the adversary successfully re-identifies a victim $u \in I$ in $t(\varphi G^+)$ is*

$$A^{S,I}_{t(\varphi G^+)}(u) = \sum_{\rho \in \Phi_{S,I}, \rho(u)=\varphi(u)} p(\rho). \tag{1}$$

In our analyses, we restrict the function p to be a probability distribution on the domain $\Phi_{S,I}$, i.e. $\sum_{\rho \in \Phi_{S,I}} p(\rho) = 1$. We also assume that p satisfies the standard *random worlds assumption* enunciated in [8,29], which expresses that, in the absence of any information in addition to $t(\varphi G^+)$, any two isomorphic subgraphs in $t(\varphi G^+)$ are indistinguishable for the adversary. We enunciate the random worlds assumption next, adapting the terminology to the one used in this paper.

Assumption 1 (*Random worlds assumption* [8,29]) Let $G = (V, E)$ be an original graph, $G^+ = (V \cup S, E')$ the sybil-extended graph created by an adversary that targets a set of victims $I \subseteq V$, and $G' = t(\varphi G^+)$ the anonymised version of G^+ created by the defender. Let ρ_1 and ρ_2 be two bijective functions from $S \cup I$ to the set of vertices $V_{G'}$ in G'. Let $G'_{\rho_1 S, \rho_1 I}$ and $G'_{\rho_2 S, \rho_2 I}$ be the two attacker subgraphs in G' that correspond to the adversary's guesses ρ_1 and ρ_2, respectively. If $G'_{\rho_1 S, \rho_1 I}$ and $G'_{\rho_2 S, \rho_2 I}$ are isomorphic, then $p(\rho_1) = p(\rho_2)$.

In the remainder of this article, we will analyse the effectiveness of various anonymisation procedures by calculating the success probability of the adversary based on Proposition 1, and we will often resort to Assumption 1 when reasoning about the adversary's belief ρ.

4 Applicability of current privacy properties against active attacks

In this section, we make, to the best of our knowledge, the first connection between passive and active attacks by formally proving that k-symmetry provides protection against active attacks. We also prove that k-symmetry is incomplete, just like (k, ℓ)-anonymity, in the sense that none of them characterises all anonymous graphs with respect to active attacks. Last, but not least, we show that neither k-symmetry implies (k, ℓ)-anonymity, nor the other way around.

4.1 *k*-symmetry: an effective privacy model against active attacks

We use the introduced privacy model to prove that k-symmetry, the strongest notion of anonymity against passive attacks, also protects against active attacks.

Definition 2 (*k-symmetry* [50]) Let Γ_G be the universe of automorphisms in G. Two vertices u and v in G are said to be automorphically equivalent, denoted $u \cong v$, if there exists an automorphism $\gamma \in \Gamma_G$ such that $\gamma(u) = v$. Because the relation $\cong$ is an equivalence relation in the set of vertices of G, let $[u]_{\cong}$ be the equivalence class of u. G is said to satisfy k-symmetry if for every vertex u it holds that $|[u]_{\cong}| \geq k$.

Theorem 1 *Let $G' = (V', E')$ be an original graph, $G^+ = (V' \cup S, E')$ the sybil-extended graph created by an adversary that targets a set of victims $I \subseteq V'$, and $t(\varphi G^+) = (V, E)$ the anonymised version of G^+ created by the defender. If $t(\varphi G^+)$ satisfies k-symmetry, then for every vertex $u \in I$ the probability of the adversary guessing the output of $\varphi(u)$ is lower than or equal to $1/k$.*

Proof Let G be a shorthand notation for $t(\varphi G^+)$. Let $\Phi_{S,I}$ be the universe of mappings from the set of vertices in $S \cup I$ to the set of vertices in G. We define a relation $\sim$ between adversary's guesses in $\Phi_{S,I}$ by

$$\rho \sim \rho' \iff G_{\rho S, \rho I} \simeq G_{\rho' S, \rho' I}$$

Because $\simeq$ is an equivalence relation, it follows that $\sim$ is also an equivalence relation. We use $\Phi_{S,I}/\sim$ to denote the partition of $\Phi_{S,I}$ into the set of equivalence classes with respect to $\sim$, and $[\rho]_\sim$ to denote the equivalence class containing ρ. Consider, given a victim u, a successful adversary guess $\rho_0 \in \Phi_{S,I}$, i.e. a mapping satisfying that $\rho_0(u) = \varphi(u)$. Our first proof step is about showing that there exist $k-1$ other mappings $\rho_1, \ldots, \rho_{k-1}$ in $[\rho]_\sim$ satisfying that

$$\forall i, j \in \{0, \ldots, k-1\}\colon i \neq j \implies \rho_i(u) \neq \rho_j(u). \tag{2}$$

Let $\rho_0(u) = v$. Because G satisfies k-symmetry, it follows that there exist $k-1$ different vertices $\{v_1, \ldots, v_{k-1}\}$ that are automorphically equivalent to v. That is to say, there exist $k-1$ automorphisms $\gamma_1, \ldots, \gamma_{k-1}$ in Γ_G such that $\forall i \in \{1, \ldots, k-1\}\colon \gamma_i(v) = v_i \neq v$. Now, consider the mappings $\rho_i \colon S \cup I \to S_i \cup I_i$ defined by $\rho_i = \gamma_i \circ \rho_0$, for every $i \in \{1, \ldots, k-1\}$. On the one hand, given that $\gamma_1, \cdots, \gamma_{k-1}$ are automorphisms, it follows that $G_{S_0, I_0} \simeq_{\gamma_i} G_{S_i, I_i}$, for every $i \in \{1, \ldots, k-1\}$, which implies that $\rho_0 \sim \rho_i$. On the other hand, $\rho_i(u) = u_i \neq u_j = \rho_j(u)$ for every $i \neq j \in \{0, \ldots, k-1\}$. This allows us to conclude that $\rho_0, \ldots, \rho_{k-1}$ are pairwise different and that $\{\rho_0, \ldots, \rho_{k-1}\} \subseteq [\rho]_\sim$.

Our second proof step consists of showing that, given two mappings ρ_0 and ρ'_0 in $\Phi_{S,I}$ such that $\rho_0(u) = \rho'_0(u) = v$, and the mappings $\{\rho_1, \ldots, \rho_{k-1}\}$ and $\{\rho'_1, \ldots, \rho'_{k-1}\}$ constructed as previously, it holds that

$$\rho_0 \neq \rho'_0 \implies \rho_i \neq \rho'_j \ \forall i, j \in \{1, \ldots, k-1\}.$$

Let $x \in S \cup I$ such that $\rho_0(x) \neq \rho'_0(x)$. Take any two integers $i, j \in \{1, \ldots, k-1\}$. We analyse two cases.

Case 1 ($i = j$). Let $\rho_0(x) = y$ and $\rho'_0(x) = y'$. By construction, $\rho_i(x) = \gamma_i(\rho_0(x)) = \gamma_i(y)$ and $\rho'_i(x) = \gamma_i(\rho'_0(x)) = \gamma_i(y')$. The fact that γ_i is a bijective function and that $y \neq y'$ gives that $\gamma_i(y) \neq \gamma_i(y')$, which implies that $\rho_i \neq \rho'_i$.

Case 2 ($i \neq j$). Observe that $\rho_i(u) = \gamma_i(\rho_0(u)) = \gamma_i(v) = v_i$ and $\rho'_j(u) = \gamma_j(\rho'_0(u)) = \gamma_j(v) = v_j$. Because $v_i \neq v_j$ it follows that $\rho_i(u) \neq \rho'_j(u)$, hence $\rho_i \neq \rho'_j$.

The last proof step consists of using the formula to calculate adversary success to obtain a probability bound. The adversary's probability of success in re-identifying a victim $u \in I$ is calculated by,

$$\sum_{\rho \in \Phi_{S,I}, \rho(u) = \varphi(u)} p(\rho).$$

Let $\rho_1^0, \ldots, \rho_n^0$ be all functions in $\Phi_{S,I}$ satisfying $\rho_1^0(u) = \rho_2^0(u) = \cdots = \rho_n^0(u) = \varphi(u)$. It follows that the probability of success of the adversary is equal to $p(\rho_1^0) + \cdots + p(\rho_n^0)$. Now, for each ρ_i, consider the mappings $\rho_i^1, \ldots, \rho_i^{k-1}$ defined by $\rho_i^j = \varphi_j \circ \rho_i^0$, for every $j \in \{1, \ldots, k-1\}$. Previously we proved the following two intermediate results.

1. For every $i \in \{1, \ldots, n\}$, the set $\{\rho_i^0, \rho_i^1, \ldots, \rho_i^{k-1}\} \subseteq \Phi_{S,I}$ has cardinality k and its elements satisfy $\rho_i^0 \sim \rho_i^1 \sim \ldots \sim \rho_i^{k-1}$.
2. $\forall i, j \in \{1, \ldots, n\} \implies \{\rho_i^0, \rho_i^1, \ldots, \rho_i^{k-1}\} \cap \{\rho_j^0, \rho_j^1, \ldots, \rho_j^{k-1}\} = \emptyset$.

The second result and the fact that p is a probability distribution give,

$$\sum_{i \in \{1,\ldots,n\}} \sum_{j \in \{0,\ldots,k-1\}} p(\rho_i^j) \leq 1.$$

We use the first result and the random worlds assumption (Proposition 1) to conclude that $p(\rho_i^0) = p(\rho_i^1) = \cdots = p(\rho_i^{k-1})$, for every $i \in \{1, \ldots, n\}$, which gives,

$$\sum_{i \in \{1,\ldots,n\}} \sum_{j \in \{0,\ldots,k-1\}} p(\rho_i^j) = \sum_{i \in \{1,\ldots,n\}} kp(\rho_i^0) \leq 1.$$

The last inequality states that $p(\rho_1^0) + \cdots + p(\rho_n^0) \leq 1/k$. □

4.2 *k*-symmetry *versus* (k, ℓ)-anonymity

As proven in Theorem 1, k-symmetry provides protection against active attacks regardless of the number of sybil nodes inserted by the attacker, as opposed to (k, ℓ)-anonymity which uses ℓ as a parameter on the maximum number of sybil nodes. In spite of that, (k, ℓ)-anonymity is not weaker than k-symmetry. As we prove next, they are in fact incomparable.

Theorem 2 *Let $\mathcal{G}_{k,\ell}$ be the universe of anonymised graphs such that no adversary with ℓ sybil nodes or less can re-identify a victim with probability lower or equal than $1/k$. There exist $k > 1$ and graphs $G, G', G'' \in \mathcal{G}_{k,\ell}$ such that:*

- *G satisfies k-symmetry, but G does not satisfy (k, ℓ)-anonymity for some $\ell \geq 1$.*
- *G' satisfies (k, ℓ)-anonymity for some $\ell \geq 1$, but G' does not satisfy k-symmetry.*
- *G'' neither satisfy k-symmetry nor (k, ℓ)-anonymity for some $\ell \geq 1$.*

Proof Figure 3a shows a 2-symmetric graph G which, for $2 \leq \ell \leq 8$, does not satisfy (k, ℓ)-anonymity for any $k > 1$. Moreover, Fig. 3b shows a $(2, 1)$-anonymous graph G' which can be verified not to satisfy k-symmetry for any $k > 1$. In fact, this graph even fails to satisfy k-degree anonymity for any $k > 1$. An example of a graph G'' proving the correctness of the last statement is displayed in Fig. 3c. That graph is neither 2-symmetric nor $(2, 2)$-anonymous. □

Of independent interest is our proof that k-automorphism [54] does not protect against active attacks. This is a surprising result, given that k-automorphism and k-symmetry have traditionally been considered equivalent. We refer the interested reader to Appendix A.

5 Algorithm K-MATCH guarantees *k*-symmetry

In this section we prove that the algorithm K- MATCH, proposed in [54] as a sufficient condition to achieve k-automorphism, also guarantees k-symmetry. Given a graph G and a

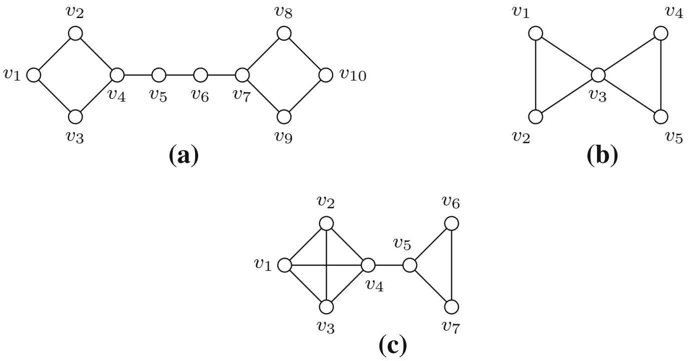


Fig. 3 Example graphs. **a** A 2-symmetric graph not satisfying (k, ℓ)-anonymity for $k > 1$ and $2 \leq \ell \leq 8$. **b** A (2, 1)-anonymous graph not satisfying k-symmetry for $k > 1$. **c** A graph where the success probability of any active attack leveraging 2 sybil nodes is at most $1/2$, despite the graph neither satisfying (2, 2)-anonymity nor 2-symmetry

value of k, the K- MATCH algorithm obtains a supergraph G' of G satisfying the following conditions:

1. $V_{G'} \supseteq V_G$ and $E_{G'} \supseteq E_G$.
2. There exist $k-1$ automorphisms $\gamma_1, \gamma_2, \ldots, \gamma_{k-1}$ of G' such that:
   (a) For every $v \in V_{G'}$ and every $i \in \{1, \ldots, k-1\}$, $\gamma_i(v) \neq v$.
   (b) For every $v \in V_{G'}$ and every $i, j \in \{1, \ldots, k-1\}$, $i \neq j \iff \gamma_i(v) \neq \gamma_j(v)$.
   (c) For every $v \in V_{G'}$ and every i, j such that $1 \leq i < j \leq k-1$, $\gamma_{i+j}(v) = \gamma_i(\gamma_j(v)) = \gamma_j(\gamma_i(v))$, with addition taken modulo k.

To obtain G', the algorithm first splits the vertices of G' into k groups and arranges them in a k-column matrix M called the *vertex alignment table* (*VAT* for short). If $|V_G|$ is not a multiple of k, a number of dummy vertices are added to achieve this property. The VAT is organised in such a manner that the number of graph editions to perform in the second step of the process is close to the minimum. For convenience, in what follows we will denote by v_{ij} the vertex of G' placed in position M_{ij} of the VAT. The second step of the method consists in adding edges to $E_{G'}$ in such a way that conditions 2.a to 2.c are enforced. To that end, for every edge (v_{ij}, v_{pq}), all edges of the form $(v_{i,j+t}, v_{p,q+t})$, additions modulo k, are added to $E_{G'}$ if they did not previously exist.

Figure 4 shows an example of a VAT allowing to enforce 3-automorphism on the graph of Fig. 2b[2]. This VAT encodes two functions $f_1, f_2\colon V_{G'} \to V_{G'}$:

$$f_1 = \{(1, F), (F, D), (D, 1), (C, A), (A, B), (B, C), (2, 3), (3, E), (E, 2)\},$$

that is, a function such that the image of every element is the one located one column to its right (modulo 3) on the same row, and

$$f_2 = \{(1, D), (F, 1), (D, F), (C, B), (A, C), (B, A), (2, E), (3, 2), (E, 3)\},$$

that is, a function such that the image of every element is the one located two columns to its right (modulo 3) on the same row.

[2] This table is not necessarily the one created by the first step of K- MATCH, but it serves to illustrate the second step, which is the one that guarantees the privacy property and will be the basis of the main result in this section.

Fig. 4 An example of a VAT for the graph shown in Fig. 2b

1	F	D
C	A	B
2	3	E

In general, these functions are not automorphisms of G' upon creation of the VAT. It is the second step of the method that will transform them into automorphisms by performing all necessary edge-copying operations. For example, the edge (C, A) needs to be added to G' because $(A, B) \in E_G$ but $(f_2(A), f_2(B)) = (C, A) \notin E_G$; and $(A, 3)$ needs to be added because $(B, E) \in E_G$ but $(f_2(B), f_2(E)) = (A, 3) \notin E_G$. Once the method is executed, each automorphism γ_t, $t \in \{1, \dots, k-1\}$, defined in item 2 above is completely specified by the VAT, as $\gamma_t(v_{ij}) = v_{i,j+t}$, with addition modulo k, for every $i \in \left\{1, \dots, \left\lceil \frac{|V_G|}{k} \right\rceil\right\}$ and every $j \in \{1, \dots, k\}$.

We now show the link between the K- Match method and k-symmetry.

Theorem 3 *Let $G = (V, E)$ be a graph and let $G' = (V', E')$ the result of applying algorithm* K- Match *to G for some parameter k. Then, G' satisfies k-symmetry.*

Proof Let $u \in V_{G'}$ be an arbitrary vertex of G', and let $v_1 = \gamma_1(u)$, $v_2 = \gamma_2(u)$, ..., $v_{k-1} = \gamma_{k-1}(u)$ be the images of u by the automorphisms $\gamma_1, \gamma_2, \dots, \gamma_{k-1}$ enforced on G' by the execution of K- Match. By definition, we have that $u \cong v_1 \cong v_2 \cong \dots \cong v_{k-1}$ and, by conditions 2.a and 2.b, they are pairwise different. Thus, $|[u]_{\cong}| = k$, hence G' is k-symmetric. □

The most relevant consequence of Theorem 3 is that algorithm K- Match can also be used for protecting graphs against active adversaries, as it will ensure that no victim is re-identified with probability greater than $1/k$.

6 Experiments

The purpose of these experiments[3] is to demonstrate the effectiveness and usability of k-symmetry, enforced using the K- Match algorithm, for protecting graphs against active adversaries leveraging a sufficiently large number of sybil nodes and the strongest attack strategy reported in the literature, namely the robust active attack introduced in [32]. Effectiveness is assessed in terms of the success rate measure used in previous works on active attacks [30–32], whereas usability is assessed in terms of several structural utility measures. In what follows, we describe the experimental setting, display the empirical results obtained and conclude the section with a discussion of these results.

6.1 Experimental setting

In order to make the results reported in this section comparable to previous works on active attacks and countermeasures against them [31,32], we study the behaviour of our

[3] We performed our experiments on the HPC platform of the University of Luxembourg [47]. In particular, we ran our experiments on the Gaia and Iris clusters of the UL HPC. Detailed descriptions of these clusters are available at https://hpc.uni.lu/systems/gaia/ and https://hpc-docs.uni.lu/systems/iris/, respectively. The implementations of the graph generators, anonymisation methods and attack simulations are available at https://github.com/rolandotr/graph.

proposed method on two collections of randomly generated synthetic graphs and two real-life datasets. For the first collection of synthetic graphs, we used Erdős–Rényi (ER) random graphs [14]. We generated 200, 000 ER graphs, 10, 000 for each density value in the set $\{0.1, 0.15, \ldots, 0.95, 1.0\}$. The second group of synthetic graphs was generated according to the Barabási–Albert (BA) model [3], which generates scale-free graphs. We used seed graphs of order 50 and every graph was grown by adding 150 vertices and performing the corresponding edge additions. The BA model has a parameter m defining the number of new edges added for every new vertex. We generated 10, 000 graphs for every value of m in the set $\{5, 10, \ldots, 50\}$. In generating each graph, the type of the seed graph was randomly selected among the following choices: a complete graph, an m-regular ring lattice, or an ER random graph of density 0.5. The probability of selecting each choice was set to $\frac{1}{3}$. In both cases, the generated synthetic graphs have 200 nodes. Based on the discussion on the plausible number of sybil nodes in Sect. 3, we make the number of sybils $\ell = \lceil \log_2 200 \rceil = 8$.

The first real-life social graph used in the experiments is the so-called *Panzarasa graph*, named after one of its creators [36]. This graph was collected from an online community of students at the University of California, Irvine. In the Panzarasa graph, a directed edge (A, B) represents that student A sent at least one message to student B. In our experiments, we used a processed version of this graph, where edge orientation, loops and isolated vertices were removed. This graph has 1, 893 vertices and 20, 296 edges. The second real-life social graph that we used was constructed from a collection of e-mail messages exchanged between students, professors and staff at Universitat Rovira i Virgili (URV), Spain [17]. For the construction of the graph, the data collectors added an edge between every pair of users that messaged each other. In doing so, they ignored group messages with more than 50 recipients. Moreover, they removed isolated vertices and connected components of order 2. The URV graph has 1, 133 vertices and 5, 451 edges. For both real-life graphs, we set the number of sybil nodes to be $\ell = \lceil \log_2 |V| \rceil = 11$.

We analyse three values for the privacy parameter k: a low value, $k = 2$; a high value, $k = 8$; and an intermediate value, $k = 5$. For every value of k, we compare the behaviour of the K- MATCH algorithm, which ensures k-symmetry, and several other anonymisation methods. We consider Mauw et al.'s algorithm for enforcing $(k, \Gamma_{G,1})$-adjacency anonymity [31], which explicitly addresses active adversaries and has demonstrated effectiveness in some instances of the active attack scenario [31,32]. Additionally, to enrich the comparison, we included perturbation methods devised in terms of other privacy notions, namely the edge-addition method proposed in [25] for enforcing k-degree anonymity (for $k \in \{2, 5, 8\}$) and the edge-set perturbation method proposed in [42] for enforcing ε-differential privacy (for $\varepsilon \in \{0.1, 0.5, 1.0\}$).

In order to build the vertex alignment table, algorithm K- MATCH requires the vertex set of the input graph to be partitioned into k subsets such that the number of edges linking vertices in different subsets is close to the minimum. We used the multilevel k-way partitioning method reported in [23], in specific its implementation included in the METIS library[4], for efficiently obtaining such a partition. The effectiveness of the anonymisation methods is measured in terms of their resistance to the robust active attack described in [32]. Thus, following the attacker–defender game described in Sect. 3, for every graph we first run the attacker subgraph creation stage. Then, for every resulting graph, we obtain all variants of anonymised graphs. Finally, for each perturbed graph, we simulate the execution of the re-identification stage and compute its success rate as defined in [32], that is

[4] Available at http://glaros.dtc.umn.edu/gkhome/views/metis.

$$SuccessRate = \begin{cases} \frac{\sum_{X \in \mathcal{X}} p_X}{|\mathcal{X}|} & \text{if } \mathcal{X} \neq \emptyset \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $\mathcal{X}$ is the set of equally-most-likely sybil subgraphs retrieved in $t(\varphi G^+)$ by the third phase of the attack, and

$$p_X = \begin{cases} \frac{1}{|\mathcal{Y}_X|} & \text{if} \quad Y \in \mathcal{Y}_X \\ 0 & \text{otherwise} \end{cases}$$

with $\mathcal{Y}_X$ containing all equally-most-likely fingerprint matchings according to X. For the collections of synthetic graphs, in order to obtain the scores used for the comparisons, we computed for every method the average of the success rates over every group of $10,000$ graphs sharing the same set of parameter choices. In the case of real-life graphs, we executed, for each perturbation method, 20 runs on the Panzarasa graph and 400 runs on the URV graph. In each of these runs, a different set of victims was randomly chosen. The final scores used for comparisons were the averaged success probabilities over every group of runs.

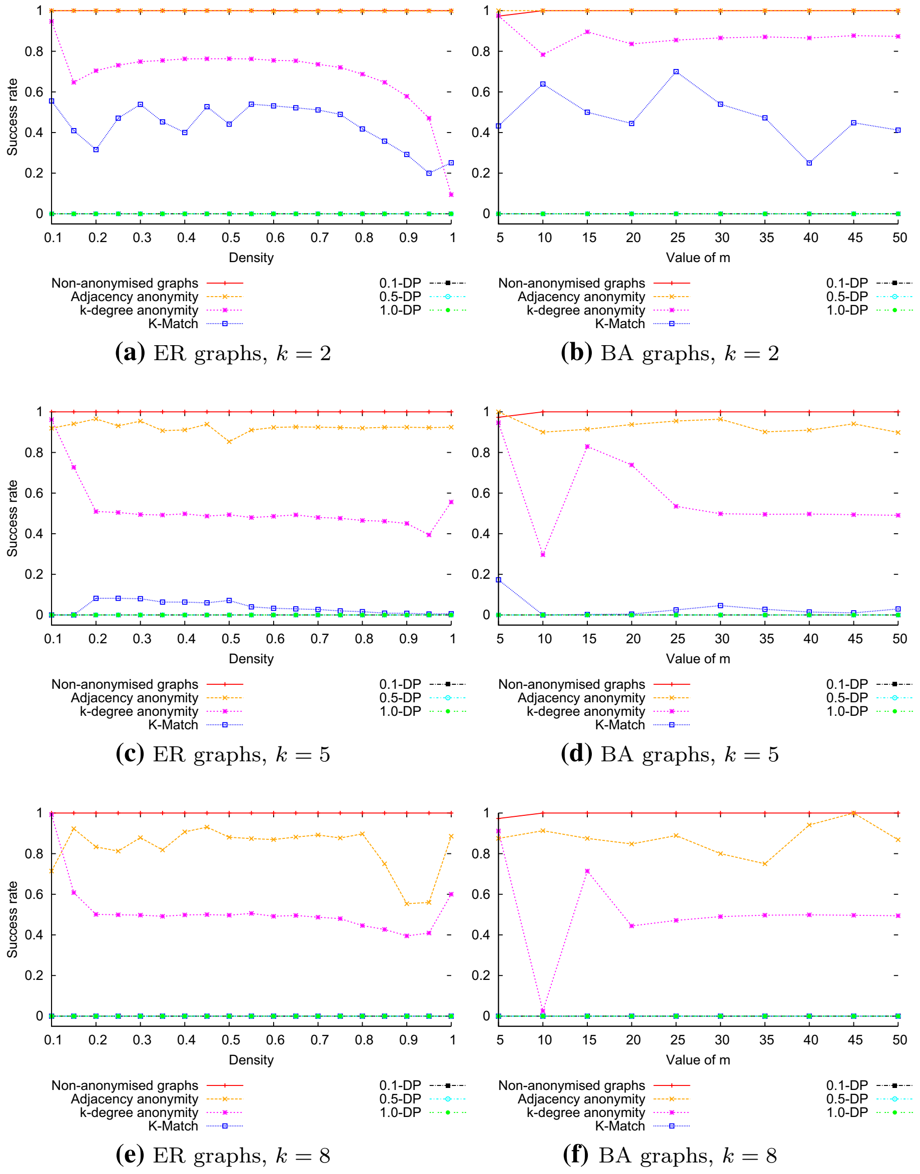
The anonymisation methods are also compared in terms of utility. To that end, we measure the distortion caused by each method on a number of global graph statistics, namely the global clustering coefficient, the averaged local clustering coefficient and the similarity between the degree distributions, measured in terms of the cosine of the angle between the degree vectors, following the approach introduced in [19,30].

6.2 Results and discussion

Figure 5 shows the success rates of the attack on both random graph collections, whereas Figs. 6, 7 and 8 show utility values in terms of degree distribution similarity, variation of global clustering coefficient and variation of averaged local clustering coefficient, respectively. Analogous results on the real-life datasets are presented in Tables 1 and 2.

Regarding the effectiveness of the anonymisation methods, the results in Fig. 5 and both tables clearly show that K- Match is considerably more effective against the robust active attack than $(k, \Gamma_{G,1})$-adjacency anonymity. These results are particularly relevant in the light of the fact that $(k, \Gamma_{G,1})$-adjacency anonymity was until now the sole formal privacy property to demonstrate non-negligible protection against the original active attack and some instances of the robust active attack [31,32]. As expected, these results show that K- Match consistently outperforms the formally weaker k-degree anonymity, displaying in most cases a significant difference. Finally, we can see that, for sufficiently large values of k, algorithm K- Match and edge-set perturbation-based differential privacy are both effective against the robust active attack. It is worth highlighting that the experiments shown here are the first ones where the robust active attack leveraging $\lceil \log_2 n \rceil$ sybil nodes is shown to be consistently thwarted by anonymisation methods based on formal privacy properties. So far, this had only been achieved in [32] via the addition of random noise, with the limitation that this work used no principled approach in determining the amount of noise to use.

Regarding utility, there is a number of scenarios where the strong protection offered by K- Match is obtained at a smaller cost than that of DP, notably for low-density and scale-free synthetic graphs, as well as both real-life graphs. Both K- Match and $(k, \Gamma_{G,1})$-adjacency anonymity have a small impact on the overall similarities of the degree distributions. This does not mean that the degrees are not affected by the methods. In fact, both methods make



(a) ER graphs, $k = 2$

(b) BA graphs, $k = 2$

(c) ER graphs, $k = 5$

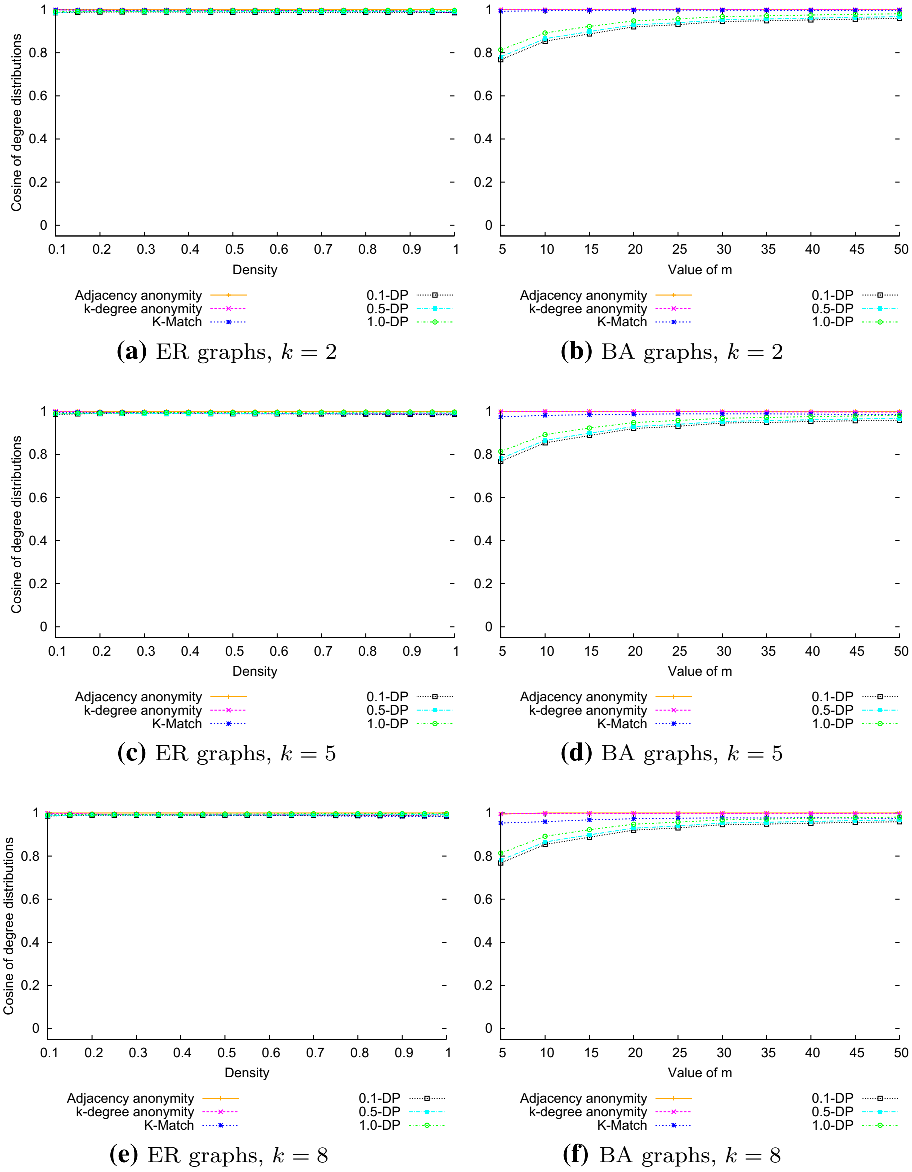
(d) BA graphs, $k = 5$

(e) ER graphs, $k = 8$

(f) BA graphs, $k = 8$

Fig. 5 Success rates of the robust active attack on the collections of Erdős–Rényi (left) and Barabási–Albert (right) random graphs, with $\ell = 8$ and $k \in \{2, 5, 8\}$

most degrees increase, but in a manner that does not significantly affect the ordering of vertices in terms of their degrees. Regarding clustering coefficient-based utilities, we can observe in Figs. 7 and 8, and both tables, that the superior effectiveness of K- Match and DP does come at the price of a larger degradation of the values of local and global clustering coefficients, although the scenarios where each method is the best differ from one method to



(a) ER graphs, $k = 2$

(b) BA graphs, $k = 2$

(c) ER graphs, $k = 5$

(d) BA graphs, $k = 5$

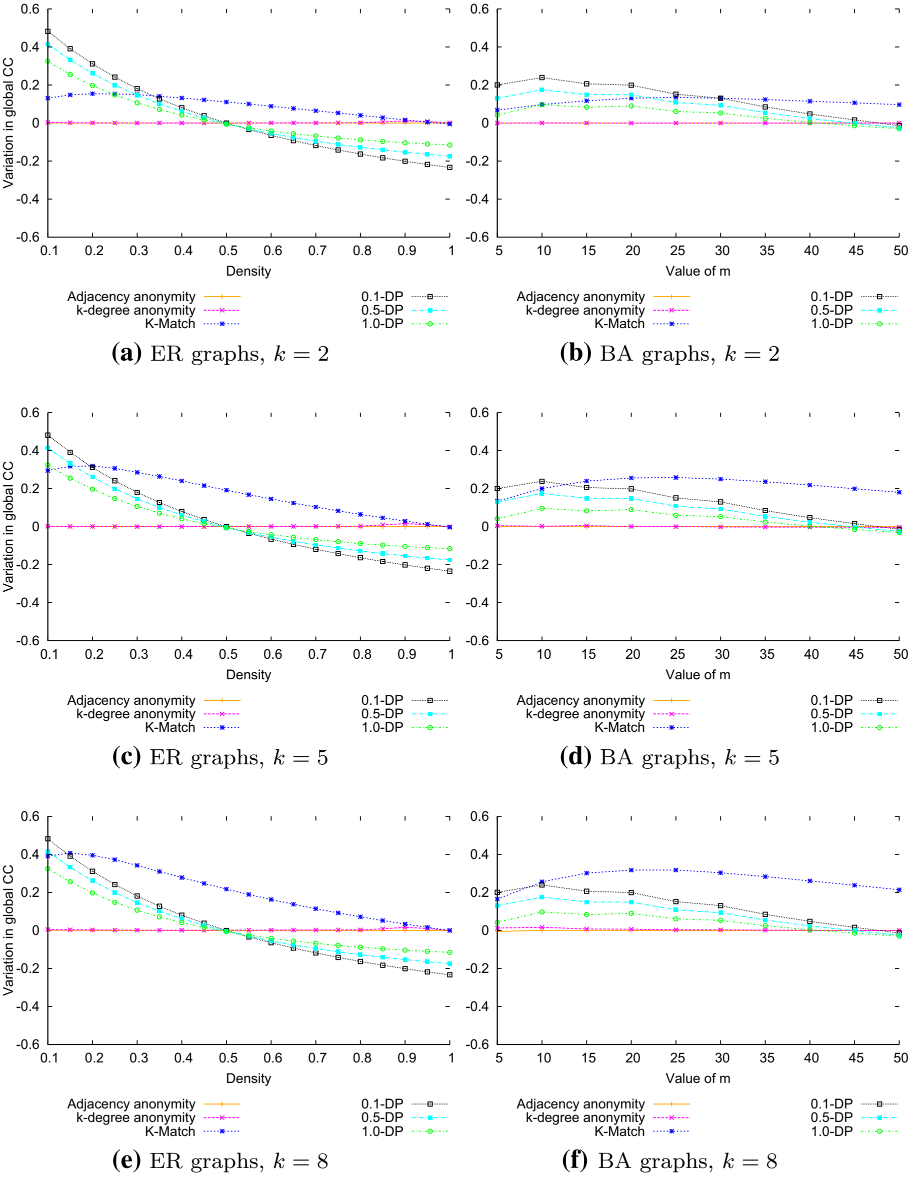
(e) ER graphs, $k = 8$

(f) BA graphs, $k = 8$

Fig. 6 Degree distribution similarities on the collections of Erdős–Rényi (left) and Barabási–Albert (right) random graphs, with $\ell = 8$ and $k \in \{2, 5, 8\}$

the other. It is worth highlighting that K- Match considerably outperforms DP in terms of most utility criteria on both real-life datasets.

In our opinion, the main takeaway from the experimental results presented in this section is that our refinement of the notion of re-identification probability for active adversaries has led to identifying, for the first time, an anonymisation method satisfying two key properties: (i) featuring a theoretically sound privacy guarantee against active attackers and (ii) having



(a) ER graphs, $k = 2$

(b) BA graphs, $k = 2$

(c) ER graphs, $k = 5$

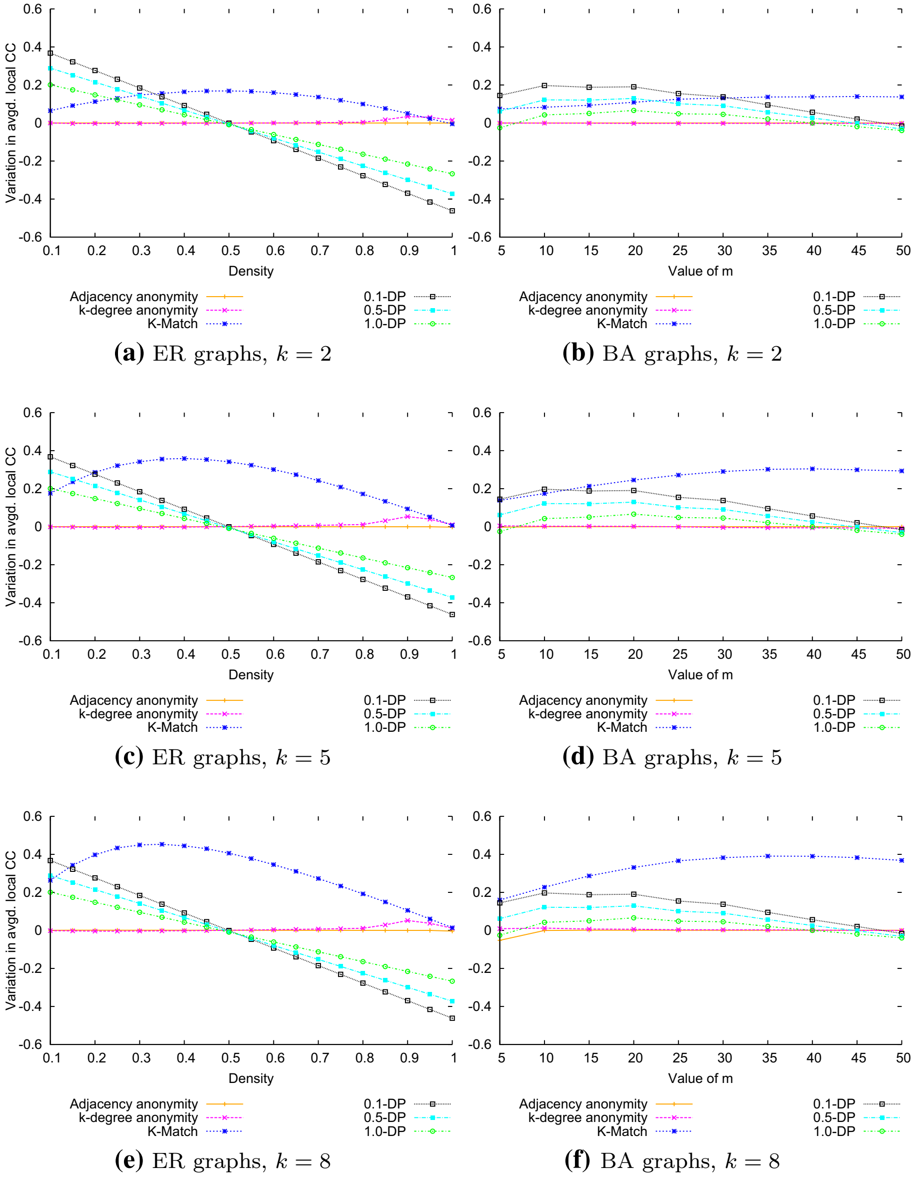
(d) BA graphs, $k = 5$

(e) ER graphs, $k = 8$

(f) BA graphs, $k = 8$

Fig. 7 Variations in global clustering coefficients on the collections of Erdős–Rényi (left) and Barabási–Albert (right) random graphs, with $\ell = 8$ and $k \in \{2, 5, 8\}$

this privacy guarantee translate into effective resistance to the strongest active attack reported so far, even when the attacker leverages a large number of sybil nodes.



(a) ER graphs, $k = 2$

(b) BA graphs, $k = 2$

(c) ER graphs, $k = 5$

(d) BA graphs, $k = 5$

(e) ER graphs, $k = 8$

(f) BA graphs, $k = 8$

Fig. 8 Variations in averaged local clustering coefficients on the collections of Erdős–Rényi (left) and Barabási–Albert (right) random graphs, with $\ell = 8$ and $k \in \{2, 5, 8\}$

7 Conclusions

We have introduced a new probabilistic interpretation of active re-identification attacks on social graphs. This enables the privacy-preserving publication of social graphs in the presence of active adversaries by jointly preventing the attacker from unambiguously retrieving the set of sybil nodes, and from using the sybil nodes for re-identifying the victims. Under the

Table 1 Results on the Panzarasa dataset

Perturb. method		Succ. rate	Sim. deg. distr.	Δ Glob. CC	Δ Av. loc. CC
None (orig. graph)		0.9984	1.0000	0.0000	0.0000
$k = 2$	k-adj. anon.	0.9986	0.9999	−0.0001	0.0016
	k-degree anon.	0.9986	0.9997	0.0046	0.0076
	K- Match	0.0041	0.9993	0.0941	0.0776
$k = 5$	k-adj. anon.	0.9984	0.9972	0.0050	0.0688
	k-degree anon.	0.9984	0.9987	0.0320	0.0190
	K- Match	0.0005	0.9967	0.1898	0.1614
$k = 8$	k-adj. anon.	0.9470	0.9896	0.0251	0.1279
	k-degree anon.	0.9987	0.9948	0.0485	0.0638
	K- Match	0.0000	0.9918	0.2278	0.1869
1.0-DP		0.0000	0.5435	0.3759	0.1608
0.5-DP		0.0000	0.5332	0.4937	0.2676
0.1-DP		0.0000	0.5287	0.5775	0.3635

Table 2 Results on the URV dataset

Perturb. method		Succ. rate	Sim. deg. distr.	Δ Glob. CC	Δ Av. loc. CC
None (orig. graph)		0.9978	1.0000	0.0000	0.0000
$k = 2$	k-adj. anon.	0.9978	0.9996	−0.0004	0.0003
	k-degree anon.	0.9978	0.9991	−0.0033	0.0012
	K- Match	0.0888	0.9991	−0.0922	−0.0824
$k = 5$	k-adj. anon.	0.9974	0.9910	0.0044	0.0208
	k-degree anon.	0.9974	0.9964	−0.0074	0.0051
	K- Match	0.0079	0.9956	−0.1080	−0.1055
$k = 8$	k-adj. anon.	0.2280	0.9665	0.0297	0.0438
	k-degree anon.	0.2326	0.9933	−0.0123	0.0085
	K- Match	0.0000	0.9890	−0.0948	−0.1055
1.0-DP		0.0000	0.7507	0.1544	0.0511
0.5-DP		0.0000	0.7435	0.2721	0.1578
0.1-DP		0.0000	0.7392	0.3559	0.2536

new formulation, we have shown that the privacy property k-symmetry provides a sufficient condition for the protection against active re-identification attacks. Moreover, we have shown that a previously existing efficient algorithm, K- Match, provides a sufficient condition for ensuring k-symmetry. Through a series of experiments, we have demonstrated that our approach allows, for the first time, to publish anonymised social graphs with formal privacy guarantees that effectively resist the robust active attack introduced in [32], which is the strongest active re-identification attack reported in the literature, even when it leverages a large number of sybil nodes.

The active adversary model addressed in this paper assumes that the (inherently dynamic) social graph is published only once. A more general scenario, where snapshots of a dynamic social network are periodically published in the presence of active adversaries, has recently been proposed in [9], and the robust active attack from [32] has been adapted to benefit from this scenario. Our main direction for future work consists in leveraging our methodology to propose anonymisation methods suited for this new publication scenario.

Acknowledgements The work reported in this paper received funding from Luxembourg's Fonds National de la Recherche (FNR), via Grant C17/IS/11685812 (PrivDA). Part of this work was conducted while Yunior Ramírez-Cruz was visiting the School of Information Technology at Deakin University.

A Appendix

It is claimed in [54] that every vertex v of a k-automorphic graph (see Definition 3) is structurally indistinguishable from $k-1$ other vertices $\varphi_1(v), \varphi_2(v), \ldots, \varphi_{k-1}(v)$.

Definition 3 (k*-automorphism* [54]) An *automorphism* is an isomorphism from a graph to itself. Formally, an automorphism γ within a graph $G = (V, E)$ is a bijective function $\gamma : V \rightarrow V$, such that $\forall v_1, v_2 \in V : (v_1, v_2) \in E \iff (\gamma(v_1), \gamma(v_2)) \in E$. A graph G is said to be k-automorphic if there exist $k-1$ non-trivial automorphisms $\varphi_1, \varphi_2, \ldots, \varphi_{k-1}$ of G such that $\varphi_i(v) \neq \varphi_j(v)$ for every $v \in V_G$ and every pair i, j satisfying $1 \leq i < j \leq k-1$.

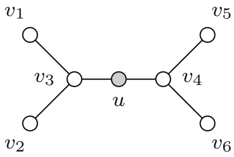


Fig. 9 A graph counterexample showing that k-automorphism does not achieve the intended privacy protection

However, a missing condition in Definition 3, namely requiring every φ_i to satisfy $\varphi_i(v) \neq v$, invalidates this claim. Consider the graph shown in Fig. 9. This graph satisfies k-automorphism as defined in Definition 3, as can be verified by the existence of the non-trivial automorphism $\gamma = \{(v_1, v_5), (v_2, v_6), (v_3, v_4), (u, u)\}$, yet the graph is vulnerable even to the simplest structural attack, the degree-based attack, as vertex u is the sole vertex with degree 2. It is worth noting that this limitation of k-automorphism does not necessarily invalidate existing anonymisation methods. This is exemplified by the K- Match algorithm itself, which does provide the intended protection because the property it directly enforces is the so-called k *different matches principle* (see [54]), which in turn is not equivalent to k-automorphism, but stronger.

References

1. Abawajy JH, Ninggal MIH, Herawan T (2016) Privacy preserving social network data publication. IEEE Commun Surveys Tutor 18(3):1974–1997

2. Backstrom L, Dwork C, Kleinberg J (2007) Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography. In: Proceedings of the 16th international conference on world wide web, pp. 181–190, New York, NY, USA
3. Barabási A, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512
4. Bonchi F, Gionis A, Tassa T (2014) Identity obfuscation in graphs through the information theoretic lens. Inf Sci 275:232–256
5. Casas-Roma J, Herrera-Joancomartí J, Torra V (2013) An algorithm for k-degree anonymity on large networks. In: Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining, pp. 671–675
6. Casas-Roma J, Herrera-Joancomartí Jordi, Torra V (2017) k-degree anonymity and edge selection: improving data utility in large networks. Knowl Inf Syst 50(2):447–474
7. Casas-Roma Jordi, Herrera-Joancomartí Jordi, Torra Vicenç (2017) A survey of graph-modification techniques for privacy-preserving on networks. Artif Intell Rev 47(3):341–366
8. Chen B-C, LeFevre K, Ramakrishnan R (2007) Privacy skyline: privacy with multidimensional adversarial knowledge. In: Proceedings of the 33rd international conference on very large data bases, VLDB '07, pp. 770–781. VLDB Endowment
9. Chen X, Këpuska E, Mauw S, Ramírez-Cruz Y (2020) Active re-identification attacks on periodically released dynamic social graphs. In *Computer Security – ESORICS 2020*, volume 12309 of *Lecture Notes in Computer Science*, pp. 185–205. Springer
10. Chen X, Mauw S, Ramírez-Cruz Y (2020) Publishing community-preserving attributed social graphs with a differential privacy guarantee. Proc Privacy Enhancing Technol 4:2020
11. Cheng J, Wai-chee FA, Liu J (2010) K-isomorphism: privacy preserving network publication against structural attacks. In: Proceedings of the 2010 ACM SIGMOD international conference on management of data, pp. 459–470
12. Chester S, Kapron BM, Ramesh G, Srivastava G, Thomo A, Venkatesh S (2013) Why Waldo befriended the dummy? k-anonymization of social networks with pseudo-nodes. Social Netw Anal Min 3(3):381–399
13. Dwork C, Roth A (2014) The algorithmic foundations of differential privacy. Found Trends Theor Comput Sci 9(3–4):211–407
14. Erdős P, Rényi A (1959) On random graphs. Publicationes Mathematicae Debrecen 6:290–297
15. Evfimievski A, Gehrke J, Srikant R (2003) Limiting privacy breaches in privacy preserving data mining. In: Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems, PODS '03, pp. 211–222, New York, NY, USA, ACM
16. Feder T, Nabar SU, Terzi E (2008) Anonymizing graphs
17. Guimera R, Danon L, Diaz-Guilera A, Giralt F, Arenas A (2003) Self-similar community structure in a network of human interactions. Phys Rev E 68(6):065103
18. Hay M, Li C, Miklau G, Jensen DD (2009) Accurate estimation of the degree distribution of private networks. In: Proceedings 19th IEEE international conference on data mining (ICDM), pp. 169–178. IEEE Computer Society
19. Ji S, Li W, Mittal P, Hu X, Beyah R (2015) Secgraph: a uniform and open-source evaluation system for graph data anonymization and de-anonymization. In: Proceedings of the 24th USENIX security symposium, pp. 303–318, Washington DC, USA
20. Jorgensen Z, Yu T, Cormode G (2016) Publishing attributed social graphs with formal privacy guarantees. In: Proceedings of the 2016 international conference on management of data, pp. 107–122
21. Karwa V, Raskhodnikova S, Smith AD, Yaroslavtsev G (2014) Private analysis of graph structure. ACM Trans Database Syst 39(3):1–33
22. Karwa V, Slavković AB (2012) Differentially private graphical degree sequences and synthetic graphs. In: Proceedings of the international conference on privacy in statistical databases, pp. 273–285
23. Karypis George, Kumar V (1998) A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM J Sci Comput 20(1):359–392
24. Li N, Li T, Venkatasubramanian S (2007) t-closeness: privacy beyond k-anonymity and l-diversity. In: Proceedings of the 23rd International conference on data engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007, pp. 106–115
25. Liu K, Terzi E (2008) Towards identity anonymization on graphs. In: Proceedings of the 2008 ACM SIGMOD international conference on management of data, pp. 93–106, New York, NY, USA
26. Lu X, Song Y, Bressan S (2012) Fast identity anonymization on graphs. In: Proceedings of the international conference on database and expert systems applications, pp. 281–295
27. Ma T, Zhang Y, Cao J, Shen J, Tang M, Tian Y, Al-Dhelaan A, Al-Rodhaan M (2015) Kdvem: a k-degree anonymity with vertex and edge modification algorithm. Computing 97(12):1165–1184
28. Machanavajjhala A, Gehrke J, Götz M (2009) Data publishing against realistic adversaries. Proc VLDB Endow 2(1):790–801

29. Martin DJ, Kifer D, Machanavajjhala A, Gehrke J, Halpern JY (2007) Worst-case background knowledge for privacy-preserving data publishing. In: Proceedings of the 23rd international conference on data engineering, ICDE 2007, The Marmara Hotel, Istanbul, Turkey, April 15-20, 2007, pp. 126–135
30. Mauw S, Ramírez-Cruz Y, Trujillo-Rasua R (2018) Anonymising social graphs in the presence of active attackers. Trans Data Privacy 11(2):169–198
31. Mauw Sjouke, Ramírez-Cruz Yunior, Trujillo-Rasua Rolando (2019) Conditional adjacency anonymity in social graphs under active attacks. Knowl Inf Syst 61(1):485–511
32. Mauw S, Ramírez-Cruz Y, Trujillo-Rasua R (2019) Robust active attacks on social graphs. Data Mining Knowl Discov 33(5):1357–1392
33. Mauw S, Trujillo-Rasua R, Xuan B (2016) Counteracting active attacks in social network graphs. In: Proceedings of DBSec 2016, vol. 9766 of *Lecture Notes in Computer Science*, pp. 233–248
34. Mir DJ, Wright RN (2009) A differentially private graph estimator. In: Proceedings 2009 ICDM international workshop on privacy aspects of data mining (ICDM), pages 122–129. IEEE Computer Society
35. Narayanan A, Shmatikov V (2009) De-anonymizing social networks. In: Proceedings of the 30th IEEE symposium on security and privacy, pp. 173–187
36. Panzarasa Pietro, Opsahl Tore, Carley Kathleen M (2009) Patterns and dynamics of users' behavior and interaction: network analysis of an online community. J Assoc Inf Sci Technol 60(5):911–932
37. Peng W, Li F, Zou X, Wu J (2012) Seed and grow: an attack against anonymized social networks. In: Proceedings of the 9th Annual IEEE communications society conference on sensor, mesh and Ad Hoc communications and networks, pp. 587–595
38. Peng Wei, Li F, Zou X, Jie W (2014) A two-stage deanonymization attack against anonymized social networks. IEEE Trans Comput 63(2):290–303
39. Rousseau F, Casas-Roma Jordi, Vazirgiannis M (2017) Community-preserving anonymization of graphs. Knowl Inf Syst 54(2):315–343
40. Sala A, Zhao X, Wilson C, Zheng H, Zhao BY (2011) Sharing graphs using differentially private graph models. In: Proceedings of the 2011 ACM SIGCOMM conference on internet measurement, pp. 81–98
41. Salas J, Torra V (2015) Graphic sequences, distances and k-degree anonymity. Discr Appl Math 188:25–31
42. Salas Julián, Torra Vicenç (2020) Differentially private graph publishing and randomized response for collaborative filtering. Proc Secrypt 2020:407–414
43. Samarati Pierangela (2001) Protecting respondents' identities in microdata release. IEEE Trans Knowl Data Eng 13(6):1010–1027
44. Stokes Klara, Torra V (2012) Reidentification and k-anonymity: a model for disclosure risk in graphs. Soft Comput 16(10):1657–1670
45. Sweeney Latanya (2002) k-anonymity: a model for protecting privacy. Int J Uncertain Fuzz Knowl Based Syst 10(5):557–570
46. Rolando T-R, Ismael GY (2016) k-metric antidimension: a privacy measure for social graphs. Inf Sci 328:403–417
47. Varrette S, Bouvry P, Cartiaux H, Georgatos F (2014) Management of an academic HPC cluster: The UL experience. In: Proceedings of the 2014 International conference on high performance computing and simulation, pp. 959–967, Bologna, Italy
48. Wang Y, Xie L, Zheng B, Lee KCK (2014) High utility k-anonymization for social network publishing. Knowl Inf Syst 41(3):697–725
49. Wang Yue, Xintao Wu (2013) Preserving differential privacy in degree-correlation based graph generation. Trans Data Privacy 6(2):127–145
50. Wu W, Xiao Y, Wang W, He Z, Wang Z (2010) K-symmetry model for identity anonymization in social networks. In: Proceedings of the 13th international conference on extending database technology, pp. 111–122
51. Xiao Q, Chen R, Tan K-L (2014) Differentially private network data release via structural inference. In: Proceedings 20th ACM SIGKDD international conference on knowledge discovery and data mining (KDD), pp. 911–920. ACM Press
52. Zhang J, Cormode G, Procopiuc CM, Srivastava D, Xiao X (2015) Private release of graph statistics using ladder functions. In: Proceedings of the 2015 ACM SIGMOD international conference on management of data, pp. 731–745
53. Zhou B, Pei J (2008) Preserving privacy in social networks against neighborhood attacks. In: Proceedings of the 2008 IEEE 24th international conference on data engineering, pp. 506–515, Washington, DC, USA
54. Zou L, Chen Lei, Tamer Özsu M (2009) K-automorphism: a general framework for privacy preserving network publication. Proc VLDB Endow 2(1):946–957

Sjouke Mauw is a Full Professor in Security and Trust of Software Systems at the University of Luxembourg. He holds a Master's degree in Mathematics and a Ph.D. in Computer Science from the University of Amsterdam. He is head of the SaToSS research group, which focuses on the application of formal methods to the design and analysis of secure systems. His research interests include security protocols, e-voting, security assessment, trust and risk management, privacy and attack trees.

Yunior Ramírez-Cruz is a postdoctoral Research Associate at the University of Luxembourg. He holds a Master's degree in Computer Science from Universidad de Oriente (Cuba) and a Ph.D. in Computer Engineering and Mathematics from Universitat Rovira i Virgili (Spain). His research interests include privacy-preserving social network analysis and publication, graph theory, data mining and the application of natural language processing to knowledge discovery, security modelling and digital forensics.

Rolando Trujillo-Rasua is a Senior Lecturer in Cyber Security at Deakin University (Australia). He completed a Master's and Ph.D. in Computer Engineering at Universitat Rovira i Virgili (Spain) and spent 5 years at the University of Luxembourg as a postdoctoral Research Associate. His research interests span the areas of formal methods, computer security and privacy protection.