k-Anonymity on Graphs using the Szemerédi Regularity Lemma

Giorgia Minello, Luca Rossi, *Member, IEEE,* and Andrea Torsello, *Member, IEEE*

Abstract—Graph anonymization aims at reducing the ability of an attacker to identify the nodes of a graph by obfuscating its structural information. In k-anonymity, this means making each node indistinguishable from at least other $k-1$ nodes. Simply stripping the nodes of a graph of their identifying label is insufficient, as with enough structural knowledge an attacker can still recover the nodes identities. We propose an algorithm to enforce k-anonymity based on the Szemerédi regularity lemma. Given a graph, we start by computing a regular partition of its nodes. The Szemerédi regularity lemma ensures that such a partition exists and that the edges between the sets of nodes behave quasi-randomly. With this partition to hand, we anonymize the graph by randomizing the edges within each set, obtaining a graph that is structurally similar to the original one yet the nodes within each set are structurally indistinguishable. Unlike other k-anonymization methods, our approach does not consider a single type of attack, but instead it aims to prevent any structure-based de-anonymization attempt. We test our framework on a wide range of real-world networks and we compare it against another simple yet widely used k-anonymization technique demonstrating the effectiveness of our approach.

Index Terms—Privacy, Anonymity, Social networks, Graph, Regularity lemma.

1 Introduction

In an era where both private companies and governments are aggressively seeking to profile users based on their online behavior, developing effective methods to anonymize datasets of users interactions before making them public is of paramount importance [14], [26], [28], [30]. These interactions are often represented using graphs that, despite potentially containing sensitive information, are made publicly available for various purposes, including research ones [6], [19], [21].

Unfortunately, a naive anonymization that simply strips the elements of a dataset of their identity (e.g., the user names) has been shown to be easily circumvented, as a combination of data attributes and external knowledge can help a malicious attacker to uniquely identify each element of the dataset [1]. Indeed, it is possible to disclose the identity of an individual participating in the network with minimal external background information. One common example is that of a user for which the number of connections in the network is known (i.e., the number of friends on Facebook) and this number happens to be unique for that individual. In other words, this piece of information alone would be sufficient to identify that user among the rest of the nodes. Most importantly, once the identity is revealed, other potentially sensitive pieces of information can be inferred. For instance, the individual may turn out to belong to a group of nodes labeled with a certain sensitive attribute, e.g., health condition.

For these reasons, the problem of anonymizing graph data is becoming an increasingly studied one [17], [22], [26], [29]. A common anonymity model is k-anonymity, which aims to ensure that each node in a network is structurally indistinguishable from at least other k nodes. Different works have focused on different definitions of "structurally indistinguishable". Liu and Terzi [22] considered the case of k-degree anonymous graphs, where k-degree anonymity guarantees that each node of the graph shares the same degree of at least k other nodes. Successive works attempted to reduce the total running time of Liu and Terzi [22] to make it feasible to scale up to large networks [17]. Rossi et al. [29], on the other hand, extended the concept of k-degree anonymity to multi-layer and time-varying graphs. Other researchers considered different structural distinguishability criteria where the attacker has increasing levels of information available to de-anonymize the nodes [5], [17], [40], however the main issue with these approaches lies in the need to add increasing amounts of noise as increasingly complex structural information needs to be obfuscated. More recently Rousseau [31] considered the problem of anonymizing a graph maximizing the amount of preserved community information. Finally, Qian et al. [26] and Ma et al. [23] looked at the complementary problem of de-anonymizing a graph in the case where the attacker has access to richer features as well as structural information.

While most of the previous k-anonymity approaches assume that the attacker has access only to a certain level of structural information (from the degree of a node, to its immediate neighborhood or even the whole graph), Foffano et al. [11] have recently proposed a k-anonymization framework where the resulting graph is not susceptible to any particular structure-based attack. Their approach is based on the Szemerédi regularity lemma [7], a well-known result of extremal graph theory. The Szemerédi regularity lemma has been successfully applied to several problems, from graph theory [18] to computer vision and pattern recognition [25], [34]. The lemma roughly states that every sufficiently large and dense graph can be approximated by the union of

- *L. Rossi is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London, United Kingdom. E-mail: luca.rossi@qmul.ac.uk*
- *G. Minello and A. Torsello are with Dipartimento di Scienze Ambientali, Informatica e Statistica, Università Ca' Foscari Venezia, Venezia, Italy.*

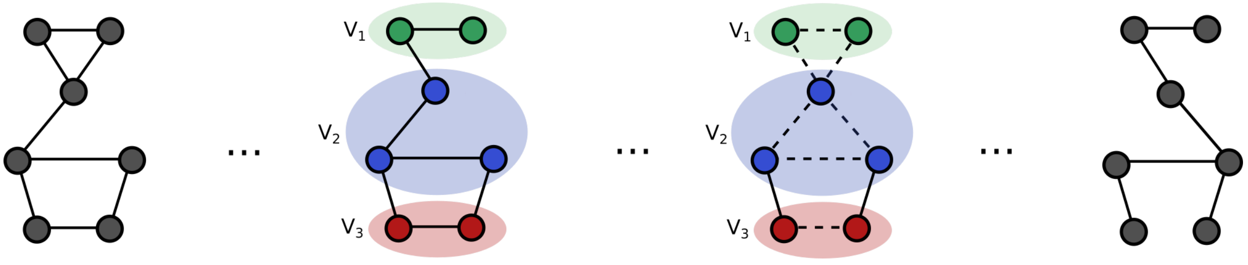


Fig. 1: Our anonymization framework consists of the following pipeline: 1) given an input graph, 2) we compute the Szemerédi partition of its nodes, and 3) we randomly re-wire the edges within each group of nodes and between each irregular pair to 4) obtain the anonymized graph. In this toy example V_1, V_2, and V_3 denote the three groups of nodes of the partition, V_i, V_j form an irregular pair, and the dashed edges denote those edges that can be added/deleted during the rewiring step.

random-like bipartite graphs called regular pairs. Crucially, the lemma was later extended to account for sparse graphs as well [15]. The idea underpinning the work of Foffano et al. [11] is that the groups of nodes that form regular pairs can be anonymized by rewiring the intra-group edges according to an Erdös-Rényi process [8]. Thanks to the theoretical guarantees of the Szemerédi regularity lemma, this has minimal effect on the overall graph structure and, together with the random-like behavior of the inter-groups connections, ensures that the each group of nodes is anonymous.

While the seminal work of Foffano et al. [11] provides a limited theoretical and experimental analysis of their method, in this paper we propose to go one step further by making the following contributions:

- we perform a theoretical analysis of the level of anonymity provided by our method;
- we devise a comprehensive experimental setting to evaluate the loss of structural information in the anonymized graphs. This involves both selecting a set of suitable metrics, as well as doing a more extensive exploration of the parameters space over a larger number of datasets;
- we compare the results obtained in this experimental setting with those achieved by another well-known k-anonymization approach, i.e., k-degree anonymity. Despite the latter being one of the least invasive anonymization approaches (as it only obfuscates local structural information, i.e., the nodes degree), we show that our method yields anonymized graphs that are significantly closer to the original ones, despite building large anonymity groups.

It should be noted that our work bears similarities with that of Hay et al. [16], who propose to anonymize graphs by randomly removing a subset of the existing edges and randomly adding a number of new edges. The idea of randomizing the structure is similar to yours, however crucially we provide a principled method to randomize only selected parts of the graph so as to try and minimize the structural information loss. Other similar works [4], [27], [35] are based on the idea of finding a partition of the nodes such that each group vertices contains at least k vertices and where the amount of noise needed to anonymize each group is minimum. The difference between these works and ours lies in the choice of the vertex partitioning algorithm, where we choose to rely on a well-studied and theoretically robust result from graph theory instead of a new heuristic.

Finally, it should be noted that in this paper we do not consider a scenario where the attacker knows the original structure of the graph as well as the node identities. In such a scenario, the attacker may be able to map the known identities from the original graph to the anonymized one, especially for small anonymization groups where the structural deviation from the original graph is minimal. With this information to hand, the attacker could then transport any node attribute that is present in the anonymized graph to the known one. However this is a different scenario from that considered in this paper, where we make the assumption that structure and identities are not available to the attacker and are instead the information to be protected. Scenarios where the attacker has access to both the original structure as well as the node identities could instead be handled adding structural noise [36] to the inter-group connections of ε-regular pairs (see Section 3).

The remainder of the paper is organized as follows. We start by introducing the key graph theoretical concepts underpinning our work in Section 2. In Section 3 we describe the anonymization method based on the Szemerédi regularity lemma and in Section 4 we evaluate our framework on six different real-world datasets. Finally, Section 5 concludes the paper.

2 Szemerédi Regularity Lemma

Let $G = (V, E)$ be an undirected graph with no self-loops, where V is the set of nodes and E is the set of edges. If V_i and V_j are two disjoint subsets of V, the edge density of the pair (V_i, V_j) is defined as

$$d(V_i, V_j) = \frac{|E(V_i, V_j)|}{|V_i||V_j|}\,, \tag{1}$$

where $E(V_i, V_j)$ is the set of edges connecting nodes in V_i to nodes in V_j. Note that the edge density satisfies $0 \leq d(V_i, V_j) \leq 1$.

Definition 2.1 (ε-regular pair). *Given a positive real number $\varepsilon > 0$, a pair of node sets V_i and V_j is called ε-regular if for all subsets $U_i \subseteq V_i$ and $U_j \subseteq V_j$ such that $|U_i| \geq \varepsilon|V_i|$ and $|U_j| \geq \varepsilon|V_j|$, we have $|d(V_i, V_j) - d(U_i, U_j)| \leq \varepsilon$.*

It follows from the above definition that the distribution of the edges between two sets forming a ε-regular pair is

TABLE 1: Summary of datasets statistics.

Dataset	Nodes	Density	Edges	Clust. Coeff.	Trans.
Twitch PT	1912	0.0171	31299	0.320	0.1309
Tv shows	3892	0.0023	17262	0.374	0.5906
Facebook	4039	0.0108	88234	0.606	0.5191
Twitch ES	4648	0.0055	59382	0.222	0.0842
Politicians	5908	0.0024	41729	0.385	0.3011
Government	7057	0.0036	89455	0.411	0.2238

TABLE 2: From partition (l) to group cardinality (k).

Dataset	$l = 4$	$l = 8$	$l = 16$	$l = 32$	$l = 64$	$l = 128$	$l = 256$
Twitch PT	478	239	120	60	30	14	7
Tv shows	973	487	243	122	61	30	15
Facebook	1010	505	252	126	63	31	15
Twitch ES	1162	581	290	145	73	36	18
Politicians	1477	739	369	185	92	46	23
Government	1764	882	441	221	110	55	27

almost uniform, i.e., the graph over $V_i \cup V_j$ behaves like a random bipartite graph. Stated otherwise, the number of edges between V_i and V_j can be seen as sampled from a binomial distribution with success probability $d(V_i, V_j)$.

Definition 2.2 (ε-regular partition)**.** *Let the node set V be divided into a partition $\mathcal{P}$ of l sets $V_0, V_1 \cdots, V_l$. $\mathcal{P}$ is an ε-regular partition if: 1) it is equitable, i.e., $|V_1| = |V_2| = \cdots = |V_l|$, 2) $|V_0| < \epsilon|V|$, and 3) all except at most εl^2 pairs (V_i, V_j) $(1 \leq i < j \leq l)$ are ε-regular.*

Note that the function of the exceptional set V_0 in the previous definition is merely technical, i.e., it allows all the other classes to have the same number of vertices. With these definitions to hand, we can finally state the following.

Lemma 2.3 (Szemerédi regularity lemma)**.** *For every positive real $\varepsilon > 0$ and every positive integer m, there exist positive integers $N = N(\varepsilon, m)$ and $M = M(\varepsilon, m)$ such that, if $G = (V, E)$ is a graph with $|V| \geq N$ nodes, there is an ε-regular equitable partition of V into l groups, where $m \leq l \leq M$.*

The Szemerédi regularity lemma states that the nodes of a graph can be grouped in such a way that the distribution of the edges between each pair of node sets is close to being random, with the exception of no more than εl^2 irregular pairs. In the words of Komlós and Simonovits [18], the lemma shows that every graph can be approximated by generalized random graphs. The implication of this is that by computing the ε-regular partition of a graph we are effectively separating structural information from noise [10].

Consider in fact a graph G and an ε-regular partition of its nodes. A reduced version of G with fewer nodes and edges can be constructed by replacing each pair of ε-regular groups with two nodes connected by an edge. As shown by the Key lemma [18], the reduced graph inherits many of the fundamental structural properties of the original graph, to the point that the graph obtained by simply replacing each pair of connected nodes of the reduced graph with a complete bipartite graph over $2t$ nodes yields a new graph that can be used as a surrogate of the original one, where $t \geq 1$ is an integer. In the next section we will show how to exploit the node partition given by the Szemerédi regularity lemma to randomize the structure in selected parts of the graph and enforce k-anonymity while minimizing the loss of structural information.

3 Anonymization Framework

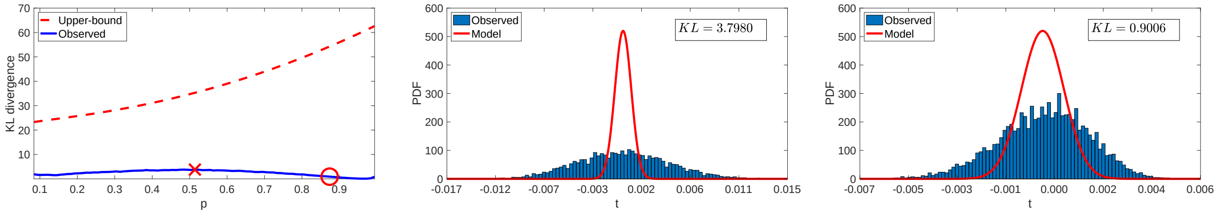
Recall that the aim of this paper is to anonymize a graph $G = (V, E)$ by grouping V into sets of k structurally indistinguishable nodes. To achieve structural anonymity within these k-anonymous groups, one needs to change the connectivity of the graph so that an attacker armed with externally acquired knowledge of the graph structure cannot distinguish among the nodes belonging to each group. For instance, rewiring all the edges of a graph with a certain probability (i.e., replacing the graph with an Erdös-Rényi graph on the same set of nodes) would yield an anonymized graph where the structural information has been rendered useless for identification purposes. However this would have come at the cost of losing most all the structural information of the original graph, thus rendering the anonymized version worthless.

Rather than achieving anonymity by disrupting the entire structure of the graph, we propose to do it selectively by making use of the Szemerédi ε-regular partition. As discussed in the previous section, the Szemerédi regularity lemma states that the node set of each graph can be partitioned to reveal a random-like structure, where pairs of groups of k nodes are randomly connected. That is, for the purpose of graph de-anonymization, the connections between two groups of nodes forming a ε-regular pair do not contain any information on the identity of the nodes belonging to the two groups as they are randomly distributed.

Other edges, on the other hand, can be still exploited to de-anonymize the nodes. These are the intra-group edges and those between a small number of irregular pairs. These are the only connections that we need to randomize in order enforce k-anonymity. Most importantly, the Szemerédi regularity lemma and the fact that the reduced graph (where the intra-group connections are lost) preserves the fundamental structural properties of the original graph imply that these intra-group connections are small in number and structurally negligible. Indeed, one way to anonymize the graph while trying to preserve its structure would be to compute its reduced version and then expand it as described in the Key lemma [18]). However, taking the particular nature of the anonymization problem into account, in this section we show that we can achieve k-anonymity in each group of the ε-regular partition without the need to lose all the intra-group structural information by passing through the reduced graph.

Our approach works as follows: 1) we first find a regular partition using the regularity lemma; 2) we randomize the groups' intra-connections; 3) finally, we randomize the edges connecting irregular pairs. More specifically:

In the **first step** we compute the ε-regular partition of the input graph using the algorithm implemented by Fiorucci et al. [9], [10]. The authors propose two different heuristic procedures where the node set is recursively split into two groups until a desired cardinality is met and the partition is sufficiently close to an ε-regular one. The two heuristics differ in the way the groups are split: 1) the *degree based* heuristic groups together nodes with similar degree, while 2) the *indeg guided* heuristic splits a sparse (dense) partition into two sparse (dense) partitions. We use the latter heuristic in our experiments as it has been shown to achieve better



(a) Upper-bound vs observed KL

(b) $p = 0.514$

(c) $p = 0.8740$

Fig. 2: Upper-bound vs observed KL divergence for varying levels of p on an ε-regular pair from the Twitch ES dataset (see Section 4). The cross and circle mark two values of p for which we show the density $f_Y(t)$ and the observed distribution of the $d(U_i, U_j)$s, respectively in (b) and (c). (b) and (c) are annotated with the corresponding value of the divergence.

results [11]. Note that due to the nature of the algorithm the cardinality of the final partition is a power of 2.

With this partition to hand, the **second step** involves randomly rewiring the connections within each group of vertices. To this end, for each group we add or delete an edge with a probability p equal to the density of the subgraph spanned by the nodes in the group. Note that we only change the internal connections of the group, so we are not altering the ε-regularity relations. The resulting subgraph has the same density of the original one, however its structural information will not be of any use when trying to de-anonymize its nodes.

The **third step** and final step is needed to randomize the connections between groups forming an ε-irregular pair. Let (V_i, V_j) be one such pair, with total number of nodes n. Consider the bipartite subgraph $H = (V_i \cup V_j, E_{ij})$ where we only consider the set of edges E_{ij} connecting nodes in V_i with nodes in V_j. In order to render the structural information contained in these edges unusable for de-anonymization purposes, we randomly rewire each pair of nodes (u, v), with $u \in V_i$ and $v \in V_j$, by adding (deleting) and edge to E_{ij} with probability p equal to $|E_{ij}|/|V_i||V_j|$.

Fig. 1 shows a toy graph as it goes through the three steps of our anonymization framework. First, the node set of the graph is partitioned into 3 groups, V_1, V_2, and V_3 (highlighted in blue, green, and red, respectively). Then edges within these groups as well as those between irregular pairs ((V_1, V_2), in this toy example) are randomly rewired while preserving the graph density. The result is an anonymized graph with 3 anonymity groups.

3.1 Quantifying the Graph Anonymity

Recall from Section 2 that given a graph $G = (V, E)$ and two disjoint sets of nodes $V_i \subseteq V$ and $V_j \subseteq V$, the density $d(V_i, V_j)$ denotes the proportion of edges connecting the nodes of V_i to the nodes of V_j. In this context, ε provides an upper bound on the absolute difference between $d(V_i, V_j)$ and $d(U_i, U_j)$, where U_i and U_j are subsets of V_i and V_j, respectively, with cardinality proportional to ε. In other words, ε is a measure of how much the ε-regular pair (V_i, V_j) deviates from a bipartite graph where the number of edges is sampled from a binomial distribution with success probability $d(V_i, V_j)$. Keeping this interpretation in mind, in this section we will show how to measure the amount of information an attacker can obtain about the identity of the nodes of a graph anonymized with our method.

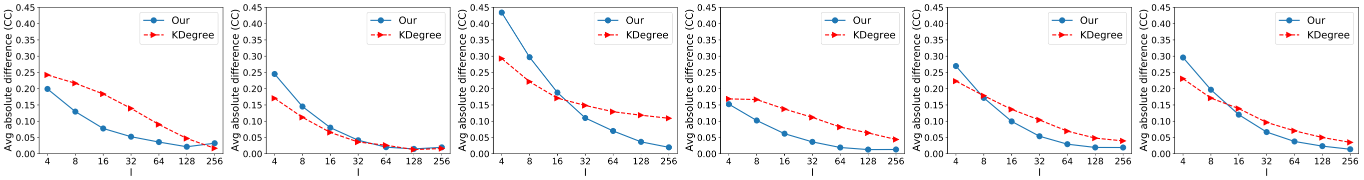
Consider a graph G and two groups forming an ε-regular pair (V_i, V_j) over k nodes each with density $\theta_{i,j} = d(V_i, V_j)$. Let U_i and U_j be two subsets of V_i and V_j respectively, each over pk nodes, where according to the definition of ε-regular partition $\varepsilon < p < 1$, i.e., p is the fraction of nodes of V_i (V_j) in U_i (U_j). Then $X \sim \text{Binomial}(p^2k^2, \theta_{V_i,V_j})$ is a Binomially distributed random variable representing the number of edges between U_i and U_j. Let $Y = \frac{X}{p^2k^2} - \theta_{i,j}$. The probability to observe a density $d(U_i, U_j) = \frac{X}{p^2k^2}$ that is at most t far from θ_{V_i,V_j} is

$$\begin{aligned} P(Y \leq t) &= P(X - p^2k^2\theta_{i,j} \leq tp^2k^2) \\ &= P\left(Z \leq \frac{tpk}{\sqrt{\theta_{i,j}(1-\theta_{i,j})}}\right), \end{aligned} \tag{2}$$

where in the last step we approximated the Binomial distribution of the random variable $Z = \frac{X - E[X]}{\sqrt{Var(X)}} = \frac{pk}{\sqrt{\theta_{i,j}(1-\theta_{i,j})}} Y$ with a standard Normal. The density function of Y is then

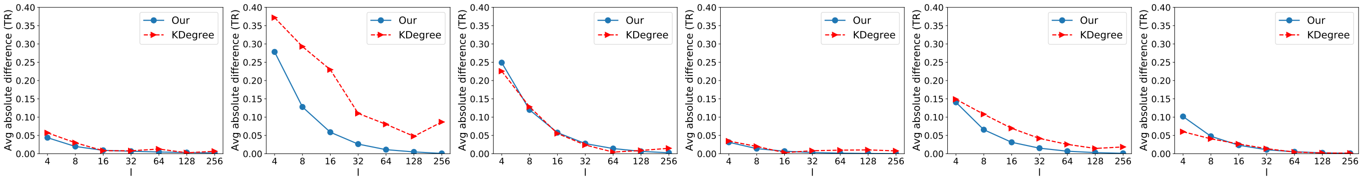
$$f_Y(t) = \frac{pk}{\sqrt{\theta_{i,j}(1-\theta_{i,j})2\pi}} e^{-\frac{1}{2}\frac{t^2p^2k^2}{\theta_{i,j}(1-\theta_{i,j})}}. \tag{3}$$

Now consider an attacker attempting to de-anonymize the nodes participating in the ε-regular pair (V_i, V_j). The degree to which the distribution of the $d(U_i, U_j)$s observed by the attacker deviates from the theoretical distribution given by Eq. 3 can be seen as a measure of the amount of structural information leaked by the anonymized graph. Note that any amount of leaked information is limited to the connections between the two groups and not necessarily all related to the nodes identity. Specifically, with Eq. 3 to hand, we measure the number of leaked bits as the Kullback-Leibler (KL) divergence between the observed distribution and the model distribution. We can give an upper bound for this quantity by considering the worst-case scenario where all the observed probability mass is equally distributed over $-\varepsilon$ and $+\varepsilon$. Recall that given two distributions p and q, their



(a) Twitch PT (b) Tv shows (c) Facebook (d) Twitch ES (e) Politician (f) Government

Fig. 3: Average absolute difference between the average clustering coefficient (ACC) of the original and anonymized graphs, for our method and k-degree.



(a) Twitch PT (b) Tv shows (c) Facebook (d) Twitch ES (e) Politician (f) Government

Fig. 4: Average absolute difference between the transitivity (TR) of the original and anonymized graphs, for our method and k-degree.

KL divergence is $KL(p||q) = \sum_x p(x) \log_2 \left(\frac{p(x)}{q(x)}\right)$. Then,

$$\begin{aligned}\text{leaked\_bits}_{V_i,V_j}(p) &= \log_2\left(\frac{1}{2}\right) - \log_2\left(\frac{f(\varepsilon)}{p^2k^2}\right)\\ &= -1 + 3\log_2(k) - \log_2(2(\theta_{i,j}(1-\theta_{i,j}))\\ &+ \log_2 p + \frac{\varepsilon^2 p^2 k^2}{2\ln(2)\theta_{i,j}(1-\theta_{i,j})}\,. \qquad (4)\end{aligned}$$

Fig. 2 (a) shows the value of the upper-bound of Eq. 4 as function of p for a regular pair computed on the Twitch ES dataset (see Section 4). Specifically, given an ε-regular partition of this graph, we chose the pair (V_i, V_j) with the highest density and we sampled $10,000$ subgroups $U_i \subseteq V_i$ and $U_j \subseteq V_j$. We then estimated the distribution of the densities $d(U_i, U_j)$ and we computed the KL divergence between this and $f_Y(t)$ (Eq. 3). The cross and circle mark two values of p for which we show the corresponding observed and model distributions.

In this example, $k = 239$ and thus approximately 8 bits are necessary to identify one node belonging to a k-anonymity group (with approximately 1888 bits required to de-anonymize the entire group). However note that the maximum value of the divergence we observe (marked with a cross) is approximately 3.8 bits. Moreover, we stress again that the bits we are measuring here refer to the information on the connections between the two groups and thus are not necessarily all useful toward identifying the nodes.

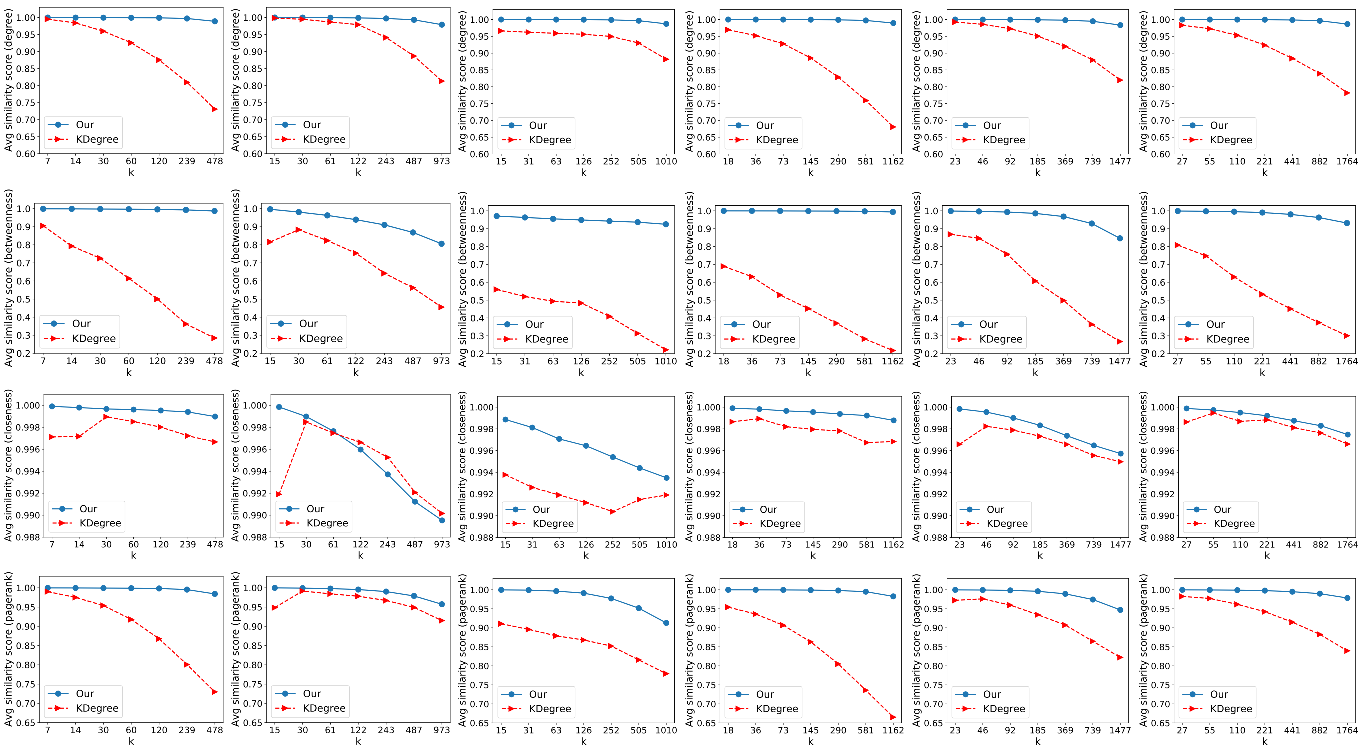
4 Experimental Evaluation

We evaluate our anonymization framework on 6 real-world networks: 1) *Twitch PT* and 2) *Twitch ES* [32] are Twitch user-user networks representing friend relations between gamers streaming in Portuguese and Spanish, respectively; 3) *Facebook* combined [20] is a network representing Facebook friend relations; 4) *TV Shows*, 5) *Politician* and 6) *Government* [33] are represent blue verified Facebook page networks of different categories, where the nodes are the pages and edges are mutual likes. Table 1 shows a summary of the structural characteristics of these 6 datasets.

For each graph, we compute the corresponding anonymized version and we measure the amount of structural information lost with respect to the original graph. More specifically, we track the changes of both network-level and vertex-level descriptors. We compute these changes for different levels of k-anonymity corresponding to different choices of the partition cardinality l. Recall in fact that in a graph with n nodes an ε-regular partition groups the vertices into l sets of cardinality $k \approx \frac{n}{l}$.

We compare our method with the well-known k-degree method [22]. Recall that k-degree only enforces anonymity at degree level, which is easily breached by an attacker with access to higher-order structural information. However for the same reason this is also one of the methods that incurs in the least amount of structural information loss. To the best of our knowledge, there is no available implementation of the k-degree method that accounts for all edge edit operations (additions, swaps, and removals). For this reason, we created our own Python implementation of the algorithm of Liu and Terzi, which is available at https://github.com/blextar/graph-k-degree-anonymity. Our implementation is based on the dynamic-programming algorithm to anonymize the degree sequence described in section 8 of [22] and the graph construction algorithm *Priority* described in section 6.2 of [22]. It should be noted that we also attempted to compare our method to the k-symmetry model of Wu et al. [39], however the only implementation of it that we could find (available at https://github.com/Keinang/K-Anonymity) was not able to cope with the large anonymity groups considered in this paper.

With both anonymization methods to hand, we proceeds as follows. For each dataset and a given value of k, we



(a) Twitch PT (b) Tv shows (c) Facebook (d) Twitch ES (e) Politician (f) Government

Fig. 5: From top to bottom, cosine similarity of: 1) degree, 2) betweenness, 3) closeness, 4) and page rank centrality.

compute 50 anonymized graphs. For a given value of k and an input graph with degree sequence d, the k-degree algorithm attempts to build a k-anonymous degree sequence $\hat{d}$. If the algorithm fails to find a graph with the required $\hat{d}$, it increments the lowest values of the original degree sequence d and then repeats the anonymization and graph construction steps. Specifically, we increment each of the 10 lowest values of the degree sequence by 1. This process is iterated until a k-anonymous graph is successfully created.

On the other hand, given a value of l our method searches for the ε-regular partition with the minimum value of ε in the range 0.01 to 0.26, with steps of 0.002. We repeat this search 10 times and we keep the ε-regular partition with the lowest value of ε. If two or more partitions share the same value of ε, we keep the one with the smallest number of irregular pairs. Finally, recall from Definition 2.2 that an ε-regular partition contains an exceptional set V_0 with cardinality at most $\varepsilon|V|$. This is used to allow the other sets to have an equal number of nodes whenever $|V|$ is not a multiple of l. By definition, the exceptional set forms an irregular pair with every other set V_i in the partition, resulting in l additional irregular pairs that need to be anonymized. We deal with this by adding enough isolated nodes to the original graph so that the total number of nodes is exactly divisible by l, thus reducing the cardinality of V_0 to zero and avoiding the need to introduce additional irregular pairs.

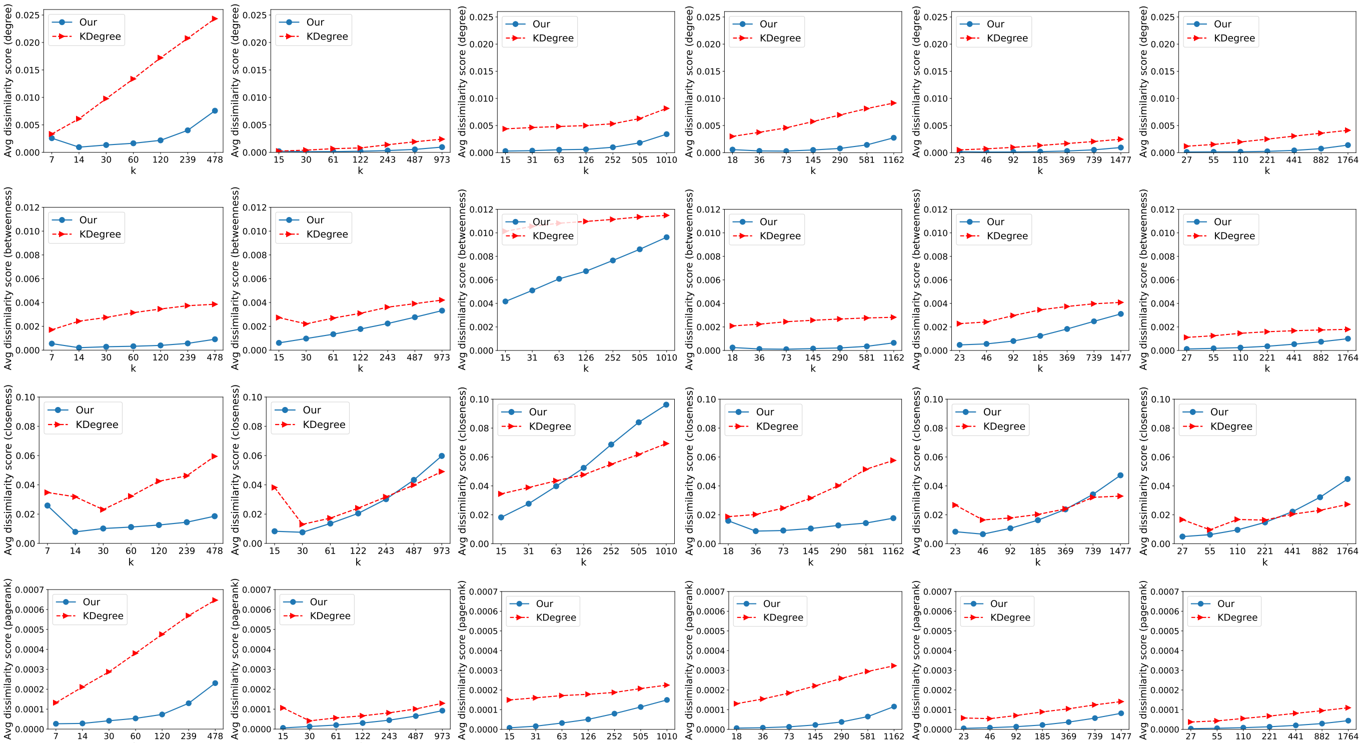
Note that our framework receives l, the value of the ε-partition cardinality we seek, in input, rather than a value of k. The value of k itself, on the other hand, depends on both the number of nodes of the graph n and the partition cardinality l. Table 2 has been included to help the reader map the values of l to the corresponding values of k on each dataset.

Finally, we want to stress that in these experiments we chose not to enforce the connectivity of the anonymized graphs produced by our method, preferring instead to let the algorithm explore a larger and less constrained optimization space in search of an optimal value of ε. A Python implementation of our method is available at https://github.com/blextar/graph-sz-anonymity.

4.1 Clustering coefficient and transitivity

We commence by characterizing the structure of both the original and the anonymized graphs using the following network-level descriptors: 1) clustering coefficient [38] and 2) transitivity [37].

Fig. 3 shows the how the absolute difference between the average clustering coefficient of the original graph and the anonymized one varies for increasing values of l (decreasing values of k). Recall that the average clustering coefficient is proportional to the number of triangles in a network. These structures in turn are likely to be broken by the Erdös-Rényi rewiring steps, particularly when the size of the anonymity group is very large. Indeed, as k decreases, less structural noise is injected in the anonymized graph and therefore the closer its clustering coefficient is to the one of the original graph. Interestingly, we note that as far as this quantity is concerned, for small values of k the k-degree method seems to fair better than ours (with the exception of the Twitch datasets). It should be stressed however that large groups,



(a) Twitch PT (b) Tv shows (c) Facebook (d) Twitch ES (e) Politician (f) Government

Fig. 6: From top to bottom, RMS of: 1) degree, 2) betweenness, 3) closeness, 4) and page rank centrality.

where our method has a clear advantage over k-degree, are preferable to small groups due to the higher level of anonymity they enforce.

Fig. 4 shows the comparison between the two methods in terms of graph transitivity, i.e., the ratio between the number of closed triplets in a graph over the maximum number of possible closed triplets, with our method performing better on the TV shows and Politicians datasets and comparably on the others.

4.2 Vertex centralities

Having focused on global network characteristics, we then look at how the vertex-level structural information of the original graphs is distorted by the anonymization procedure. Specifically, we compute the following vertex-level descriptors: 1) degree centrality, 2) betweenness centrality [3], 3) closeness centrality [13], and 4) PageRank [24]. Given two graphs G and $\tilde{G}$ and their corresponding vertex-level descriptors x and $\tilde{x}$, we compare them using the following measures: 1) cosine similarity $CS(x, \tilde{x}) = \frac{x^T \tilde{x}}{||x||||\tilde{x}||}$ and 2) Root Mean Square (RMS) $RMS(x, \tilde{x}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \tilde{x}_i)^2}$, where x_i is the value of the centrality for the vertex v_i of G and $\tilde{x}_i$ is the value of the centrality for the vertex v_i of $\tilde{G}$.

Figs. 5 and 6 show how these measures vary as we increase k, for different datasets (columns) and different centralities (rows).With some exceptions, once again reducing the value of k results in better approximations, although this comes at the expense of decreased anonymity. In all cases, however, we see that our method yields anonymized graphs that are significantly closer to the original ones in terms of vertex centrality when compared to the k-degree method. This in turns shows that our framework can generate very large anonymity groups with minimal information loss, as far as vertex centrality measures are concerned. Note that the occasional drops (increases) of the cosine similarity (RMS) for the k-degree method when k is small are due to the injection of noise in the original graph degree sequence that follow an unsuccessful graph construction given an anonymized degree sequence.

4.3 Degree distribution and edge intersection

Finally, we consider two last structural properties of the graphs, namely the degree distribution and the edge intersection between the original and the anonymized graphs, where the latter is defined as the ratio of original edges that are also in the anonymized graph [22].

Fig. 7 shows the log-log plots of the degree distributions of a selected number of graphs. In particular, we plot the degree distributions of the original graphs (green) and the graphs anonymized with our method (blue) and k-degree (red) for varying levels of k. The scatter plots show that our method is significantly better at approximating the original degree sequence. This is not surprising, in fact in order to build large anonymity groups the k-degree anonymity requires grouping together a large number of nodes to which the same degree is assigned, effectively generating a markedly different degree distribution from the original graph. This should be contrasted with our approach, which

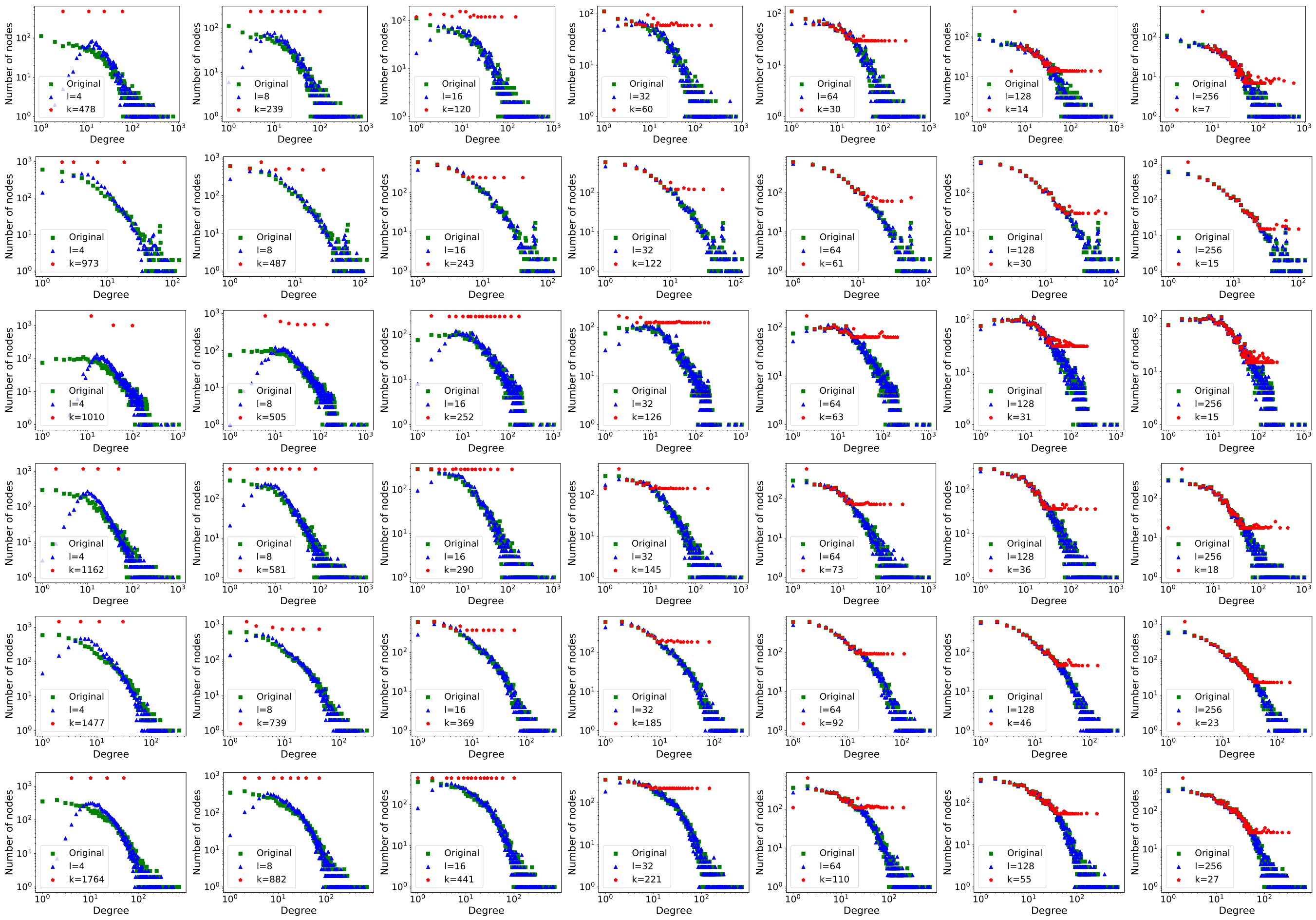


Fig. 7: Degree distributions of the original graphs (green) and graphs anonymized with our method (blue) and k-degree (red) for varying levels of k (columns) and on the 1) Twitch PT, 2) Twitch ES, 3) Politicians, 4) Facebook combined, and 5) Government datasets (rows, top to bottom).

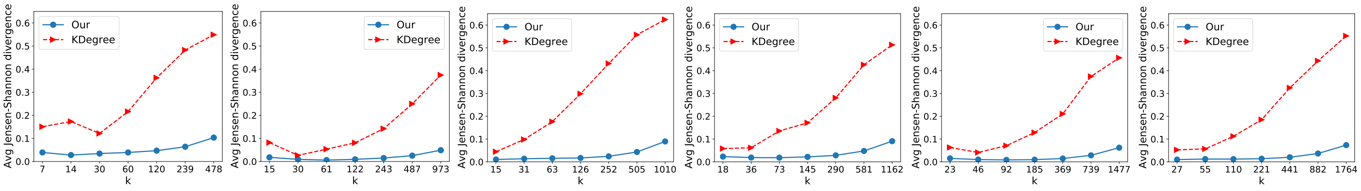
is able to create large anonymity groups while maintaining an anonymized degree distribution that closely resembles the one of the original graph.

In an attempt to better quantify the distortion of the degree distributions, we compare the distributions computed on the original graphs and the ones computed on the anonymized graphs in terms of Jensen-Shannon divergence (JSD), a well-known dissimilarity measure between probability distributions. The JSD is a symmetrized and smoothed version of the Kullback-Leibler divergence and takes values between 0 and 1. Fig. 8 shows how the JSD varies for increasing size k of the anonymity groups. Recall that a lower value of the JSD between the (normalized) degree distributions means a lower dissimilarity and thus a higher similarity between the degree sequences. Once again, our method preserves the degree information significantly better than k-degree.

Fig 9 shows instead how the edge intersection between original and anonymized graphs varies as we increase the value of k. The plot illustrates well the fact that selecting the portions of the graphs to randomize using the ε-regular partition as a guide provides an effective way to achieve anonymity while minimizing the number of edges that need to be deleted or added with respect to the original graph.

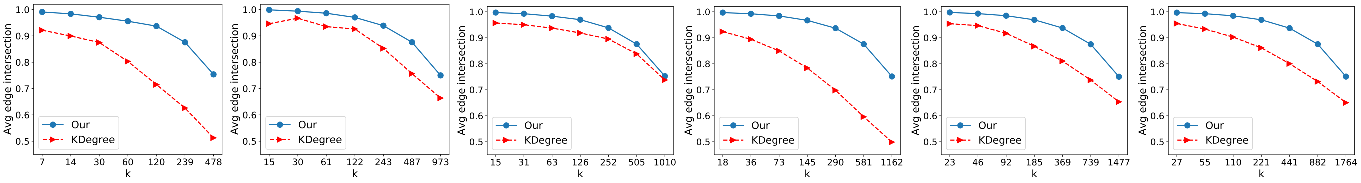
4.4 Runtime analysis

As a final experiment, we analyse the runtime of our method. Table 3 shows the average runtime (± standard error) in seconds for increasing values of l and for each dataset. For each value of l and each choice of dataset, the average is computed over the 50 realisation of the anonymized graph. All runtimes are computed on a 4 cores Xeon 5122 @ 3.60GHz with 192GB of RAM. As expected, the runtime increases monotonically with l as well as the vertex set cardinality. The only exception to this monotonic increase is observed in the Twitch PT dataset when $l = 32$. This appears to be due to the particular nature of this dataset, where a significant percentage of the nodes are hubs connected to at least 10% of the remaining nodes, as well as the difficulty of distributing them among the l groups, which results in a higher number of iterations needed to find a ε-regular partition.



(a) Twitch PT (b) Tv shows (c) Facebook (d) Twitch ES (e) Politician (f) Government

Fig. 8: The Jensen-Shannon divergence between the degree distributions. Datasets: (a) Twitch PT, (b) Tv shows, (c) Facebook combined and (d) Twitch ES.



(a) Twitch PT (b) Tv shows (c) Facebook (d) Twitch ES (e) Politician (f) Government

Fig. 9: Edge intersection between original and anonymized graphs. Datasets: (a) Twitch PT, (b) Tv shows, (c) Facebook combined and (d) Twitch ES.

TABLE 3: Average runtime (in seconds) of our method on the datasets considered in this study, for increasing values of l.

Dataset	$l = 4$	$l = 8$	$l = 16$	$l = 32$	$l = 64$	$l = 128$	$l = 256$
Twitch PT	1.28	1.85	3.91	21.58	9.99	17.57	54.84
Tv shows	8.76	12.42	16.81	37.63	48.83	67.79	106.35
Facebook	14.83	20.45	30.82	40.00	52.41	68.19	173.78
Twitch ES	19.25	26.72	36.35	51.55	61.74	93.91	190.27
Politicians	36.00	48.97	72.40	88.68	111.02	139.88	187.72
Government	49.15	76.57	111.85	158.40	199.93	247.71	312.03

5 Conclusion

We introduced a novel framework for the structural anonymization of nodes participating in a network. Our framework is based on the Szemerédi regularity lemma, a well-know theoretical result from graph theory. The key idea behind our approach is that of achieving k-anonymity by selectively randomizing certain portion of the graphs identified by the ε-regular pairs. This allows us to create anonymous groups that are resilient to any type of structural attack while minimizing the structural information loss. We validated our framework by performing an extensive set of experiments on a large number of popular real-world datasets and considering a variety of structural measures. The experimental evaluation confirmed the efficacy of our approach in generating large anonymity groups with significantly lower information loss when compared to a widely used alternative, namely k-degree anonymity.

References

[1] Backstrom, L., Dwork, C., and Kleinberg, J. (2007). Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of WWW'07* (ACM), 181–190

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

[3] Brandes, U. (2008). On variants of shortest-path betweenness centrality and their generic computation. *Social Networks* , 30(2), 136–145

[4] Campan, A. and Truta, T. M. (2008). Data and structural k-anonymity in social networks, In *International Workshop on Privacy, Security, and Trust in KDD* (Springer), 33–54

[5] Cheng, J., Fu, A. W.-c., and Liu, J. (2010). K-isomorphism: privacy preserving network publication against structural attacks. In *Proceedings of SIGMOD'10* (ACM), 459–470

[6] Chorley, M. J., Rossi, L., Tyson, G., and Williams, M. J. (2016). Pub crawling at scale: tapping untappd to explore social drinking. In *Tenth International AAAI Conference on Web and Social Media*

[7] Diestel, R. (2012). Graph theory, volume 173 of. *Graduate texts in mathematics* , 7

[8] Erdős, P. (1960). Graphs with prescribed degrees of vertices (hungarian). *Mat. Lapok* 11, 264–274

[9] Fiorucci, M., Torcinovich, A., Curado, M., Escolano, F., and Pelillo, M. (2017). On the interplay between strong regularity and graph densification. In *International Workshop on Graph-Based Representations in Pattern Recognition* (Springer), 165–174

[10] Fiorucci, M., Pelosin, F., and Pelillo, M. (2020). Separating structure from noise in large graphs using the regularity lemma, *Pattern Recognition* 98, 107070

[11] Foffano, D., Rossi, L., and Torsello, A. (2019). You Can't See Me: anonymizing Graphs using the Szemeredi Regularity Lemma. *Frontiers in Big Data*, 2, 7

[12] Freeman, L. C. (1977). A set of measures of centrality based upon betweenness. *Sociometry*, 40, 35—41

[13] Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*,1(3), 215–239

[14] Fung, B., Wang, K., Chen, R., and Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Computing Surveys* 42, 14

[15] Gerke, S. and Steger, A. (2005). The sparse regularity lemma and its applications. *Surveys in combinatorics* 327, 227–258

[16] Hay, M., Miklau, G., Jensen, D., Weis, P., and Srivastava, S. (2007). Anonymizing social networks. Computer science department faculty publication series, 180

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

[18] Komlós, J. and Simonovits, M. (1996). Szemerédi's regularity lemma and its applications in graph theory

[19] Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of WWW'10* (ACM), 591–600

[20] Leskovec, J. and Mcauley, J. J. (2012). Learning to discover social circles in ego networks. In *Advances in neural information processing systems*. 539–547

[21] Lima, A., Rossi, L., and Musolesi, M. (2014). Coding together at scale: GitHub as a collaborative social network. In *Eighth International AAAI Conference on Weblogs and Social Media*.

[22] Liu, K. and Terzi, E. (2008). Towards identity anonymization on graphs. In *Proceedings of SIGMOD'08* (ACM), 93–106

[23] Ma, J., Qiao, Y., Hu, G., Huang, Y., Sangaiah, A. K., Zhang, C., et al. (2018). De-anonymizing social networks with random forest classifier. *IEEE Access* 6, 10139–10150

[24] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66, Stanford InfoLab

[25] Pelillo, M., Elezi, I., and Fiorucci, M. (2017). Revealing structure in large graphs: Szemerédi's regularity lemma and its use in pattern recognition. *Pattern Recognition Letters* 87, 4–11

[26] Qian, J., Li, X.-Y., Zhang, C., and Chen, L. (2016). De-anonymizing social networks and inferring private attributes using knowledge graphs. In *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications* (IEEE), 1–9

[27] Ros-Martín, M., Salas, J., and Casas-Roma, J. (2019). Scalable non-deterministic clustering-based k-anonymization for rich networks. *International Journal of Information Security* 18(2), 219–238

[28] Rossi, L. and Musolesi, M. (2014). It's the way you check-in: identifying users in location-based social networks. In *Proceedings of the second ACM conference on Online social networks* (ACM), 215–226

[29] Rossi, L., Musolesi, M., and Torsello, A.. On the k-anonymization of time-varying and multi-layer social graphs. In *Ninth International AAAI Conference on Web and Social Media*

[30] Rossi, L., Williams, M., Stich, C., and Musolesi, M.. Privacy and the city: User identification and location semantics in location-based social networks. In *Ninth International AAAI Conference on Web and Social Media*

[31] Rousseau, F., Casas-Roma, J., and Vazirgiannis, M. (2018). Community-preserving anonymization of graphs. *Knowledge and Information Systems* 54, 315–343

[32] Rozemberczki, B., Allen, C., Sarkar, R., and Sutton, C. (2019). Multi-scale Attributed Node Embedding. *arXiv preprint arXiv:1909.13021*

[33] Rozemberczki, B., Allen, C., Sarkar, R., and Sutton, C. (2019). GEMSEC: Graph Embedding with Self Clustering. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2019* 65–72

[34] Sperotto, A. and Pelillo, M. (2007). Szemerédi's regularity lemma and its applications to pairwise clustering and segmentation. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition* (Springer), 13–27

[35] Stokes, K., and Torra, V. (2012). Reidentification and k-anonymity: a model for disclosure risk in graphs. *Soft Computing* 16(10), 1657–1670.

[36] Torra, V., and Salas, J. (2019). Graph Perturbation as Noise Graph Addition: A New Perspective for Graph Anonymization. In *Data Privacy Management, Cryptocurrencies and Blockchain Technology* (Springer), 121–137.

[37] Wasserman, S., and Faust, K. (1994). *Social network analysis: Methods and applications (Vol. 8)* Cambridge university press.

[38] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *Nature* 393, 440

[39] Wu, W., Xiao, Y., Wang, W., He, Z., and Wang, Z. (2010). K-symmetry model for identity anonymization in social networks. *Proceedings of the 13th international conference on extending database technology* 111–122

[40] Zhou, B. and Pei, J. (2011). The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. *Knowledge and Information Systems* 28, 47–77

Giorgia Minello Giorgia Minello worked in the industry for almost 10 years before receiving, in 2019, her PhD in Computer Science at the Ca' Foscari University of Venice (Italy), where she currently holds a Postdoctoral Research Fellow position. Her research focus is mainly on Structural Pattern Recognition and Machine Learning. The leading topics involve the analysis of networks, specifically quantum approaches to structural analysis.

Luca Rossi Luca Rossi received his PhD in Computer Science from Ca' Foscari University of Venice (Italy) in 2013. He is a Lecturer at Queen Mary University of London (UK), having held various positions at the University of Birmingham (UK), Aston University (UK), and Southern University of Science and Technology (China). He has published over 50 papers in international journals and conferences. His research interests are in the areas of pattern recognition, data mining, and network science. He is currently a member of the Editorial Board of the journal Pattern Recognition.

Andrea Torsello Andrea Torsello received his PhD in computer science at the University of York, UK. From 2007 he is with Ca' Foscari University of Venice, Italy, where he is Full Professor. His research interests are in the areas of Computer Vision and Pattern Recognition, in particular the interplay between Stochastic and Structural approaches as well as Game-Theoretic and Physical models, with applications in 3D reconstruction and recognition. Recently he has focussed on the application of Structural Pattern Recognition techniques to Network Science. Professor Torsello has published over 150 technical papers in refereed journals and conference proceedings and is repeatedly in the program committees of various international conferences and workshops reference for the area. He is currently a member of the Editorial Boards of the international journals Pattern Recognition and Pattern Recognition Letters.