

Using Metrics Suites to Improve the Measurement of Privacy in Graphs

Yuchen Zhao and Isabel Wagner, *Senior Member, IEEE*

Abstract—Social graphs are widely used in research (e.g., epidemiology) and business (e.g., recommender systems). However, sharing these graphs poses privacy risks because they contain sensitive information about individuals. Graph anonymization techniques aim to protect individual users in a graph, while graph de-anonymization aims to re-identify users. The effectiveness of anonymization and de-anonymization algorithms is usually evaluated with privacy metrics. However, it is unclear how strong existing privacy metrics are when they are used in graph privacy. In this paper, we study 26 privacy metrics for graph anonymization and de-anonymization and evaluate their strength in terms of three criteria: *monotonicity* indicates whether the metric indicates lower privacy for stronger adversaries; for within-scenario comparisons, *evenness* indicates whether metric values are spread evenly; and for between-scenario comparisons, *shared value range* indicates whether metrics use a consistent value range across scenarios. Our extensive experiments indicate that no single metric fulfills all three criteria perfectly. We therefore use methods from multi-criteria decision analysis to aggregate multiple metrics in a metrics suite, and we show that these metrics suites improve monotonicity compared to the best individual metric. This important result enables more monotonic, and thus more accurate, evaluations of new graph anonymization and de-anonymization algorithms.

Index Terms—graph anonymization, graph de-anonymization, privacy, privacy metrics, monotonicity, metrics suites

1 INTRODUCTION

The usage of Internet-based communications systems such as email or social networks leaves traces that can be collected and stored in graph form. In these graphs, nodes represent users and edges represent relationships between users. Many graph data sets have already been published for scientific or commercial use [13], [14]. Graph data can help us understand social networks [12] and improve recommendations [16], but can also harm user privacy because of sensitive information revealed by relationships.

To protect privacy, graphs can be anonymized by removing node identifiers and by changing the graph structure. As a result, the nodes in the anonymized graph cannot easily be mapped to their original identifiers if the adversary does not possess additional knowledge. However, a common assumption is that adversaries know about an auxiliary graph with similar structure as well as the correct mapping for a small number of nodes. Finding new methods for anonymization and de-anonymization is an active research area [7], and researchers usually use privacy metrics to evaluate the effectiveness of their new methods.

The most commonly used metric in graph privacy is the *adversary’s success rate*, which gives the percentage of correctly re-identified nodes [8], [20], [23]. However, even though we show in this paper that the adversary’s success rate is indeed a good metric in many scenarios, it has two important shortcomings: First, it does not reveal much detail about the privacy of individual nodes because it measures

privacy on a per-graph level. Second, its common definition favors de-anonymization algorithms that primarily rely on local, instead of global, graph properties.

To find a better metric for graph privacy, we analyze privacy metrics proposed in other fields [32] in terms of three criteria: monotonicity, evenness, and shared value range [36]. Monotonicity requires that privacy metrics indicate lower privacy as the adversary’s strength increases. Evenness requires that the metric values are spread evenly over their value range, which improves within-scenario comparisons. Shared value range requires that metrics use a common value range even when they are applied to different datasets, anonymization, or de-anonymization algorithms. This improves between-scenario comparisons.

In this paper, we make the following contributions in the area of privacy measurement:

- We propose a framework for the evaluation of privacy metrics for graph privacy based on our previous methodology [30], [36].
- We conduct extensive experiments to analyze the strength of 26 privacy metrics using 11 graph datasets, 6 anonymization algorithms, and 6 de-anonymization algorithms and find that no single metric excels in all three criteria.
- We find that several popular metrics are not monotonic in graph privacy, including entropy and the anonymity set size, and give a detailed analysis why this is the case. This finding is in contrast to results for other domains, such as vehicular network privacy [31].
- We propose four concrete metrics suites to improve the measurement of graph privacy and synthesize the metrics using the Weighted Product Model from

• Y. Zhao is with the School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, United Kingdom.
 • I. Wagner (corresponding author) is with the Cyber Security Centre, De Montfort University, Leicester, LE1 9BH, United Kingdom.
 E-mail: yuchen.zhao@soton.ac.uk, isabel.wagner@dmu.ac.uk

Manuscript received MM DD, YYYY.

multi-criteria decision analysis. We show that our metrics suites have higher monotonicity than the best individual metric.

Our findings are important because our synthesis of metrics suites shows a convenient way how the strengths of multiple privacy metrics can be combined to improve the overall measurement of privacy.

2 RELATED WORK

2.1 Graph Anonymization and De-anonymization

Structural graph de-anonymization algorithms and corresponding anonymization algorithms have been an active research area since the mid-2000s [2], when it was discovered that simply removing identifiers from nodes is not sufficient to prevent node re-identification. Since then, many anonymization and de-anonymization algorithms have been proposed [7].

In this paper, we use existing algorithms and an existing implementation [8] to evaluate the strength of privacy metrics for graph anonymization and de-anonymization. We give more detail about the algorithms we used in Sections 3.2 and 3.3.

2.2 Privacy Metrics for Graph Privacy

The most commonly used privacy metrics in graph anonymization and de-anonymization are the number of re-identified nodes [19] and the adversary's success rate [8], [20], [23]. These metrics quantify privacy as the actual privacy breach that an adversary causes.

In the wider privacy research area, other classes of privacy metrics have been proposed [32], for example measuring the adversary's uncertainty, information gain, or error. For example, information-theoretic metrics such as entropy quantify privacy as the uncertainty that the adversary faces when re-identifying nodes. These metrics have shown good strength in other fields such as vehicular networks [36], but whether they are suitable for graph privacy remains unknown. In this paper, we examine the suitability and strength of a wide range of privacy metrics in graph privacy.

De-anonymizability quantification focuses on quantifying structural properties of graph pairs, such as the edge difference, to study the maximum number of nodes that could be de-anonymized, given only structural graph information for a graph and a partially overlapping auxiliary graph [6], [21]. However, de-anonymizability quantification does not consider the interplay between anonymization and de-anonymization algorithms and has limited applicability in concrete practical scenarios due to its focus on theoretical limits for abstract graph models [7].

2.3 Criteria for Privacy Metrics

Most privacy metrics do not meet the mathematical criteria for metrics (non-negativity, identity of indiscernibles, symmetry, and triangle inequality). Instead, several authors have proposed other criteria that good privacy metrics should fulfill. For example, metrics should show the adversary's chances of success [1], show the potential for privacy violations [3], measure the amount of resources needed by

the adversary [26], and integrate measurements for different aspects of privacy [24]. However, these criteria do not allow to directly compare the strength of different privacy metrics. To this end, we have previously proposed that privacy metrics should be monotonic, i.e. that they should show lower privacy levels for stronger adversaries [30], [36].

Most of the work that compares different privacy metrics and analyzes in which situations privacy metrics perform well is in anonymous communication. For example, Syverson examines entropy as a metric for anonymity and concludes that it should not be used in the context of anonymous communication [26], and Murdoch compares the strengths and weaknesses of several different metrics for anonymous communication [18].

In our previous work, we have proposed a methodology to systematically evaluate different privacy metrics based on the criterion of monotonicity, and we have applied this methodology to privacy metrics for genomic privacy [30] and vehicular networks [31], [36]. In this paper, we adapt our methodology to examine metrics in graph privacy.

3 METHODOLOGY

We evaluate the strength of privacy metrics for social graphs in terms of monotonicity, evenness, and shared value range. In this section, we explain our methodology, which we have adapted from our prior work [30], [36], [37]. Our method follows six steps, as illustrated in Figure 1:

- 1) Import an existing graph data set (Section 3.1)
- 2) Anonymize the input graph with different anonymization algorithms (Section 3.2)
- 3) Subset the input graph to create an auxiliary graph and define seed mappings as prior information given to the adversary (Section 3.3)
- 4) Apply different de-anonymization algorithms to the anonymized graphs, controlling the adversary's strength using the amount of prior information given to the adversary (Section 3.3)
- 5) Compute the values of different privacy metrics based on the output of the de-anonymizer (Section 3.4)
- 6) Analyze the strength of each privacy metric according to the criteria of monotonicity (Section 3.5), evenness, and shared value range (Section 3.6)

3.1 Graph Datasets

We used eleven graph datasets, all available from Konect [13], to evaluate the strength of privacy metrics. Table 1 summarizes the datasets and highlights some of the graph statistics commonly used to characterize graphs. Our selection represents a diverse sample of graphs with respect to 19 graph statistics available in Konect.

3.2 Anonymizing Algorithms

Graph anonymization algorithms modify the nodes or edges of a graph to prevent the individual nodes from being identified. The simplest method is to remove the original identifiers of all nodes while leaving the graph structure

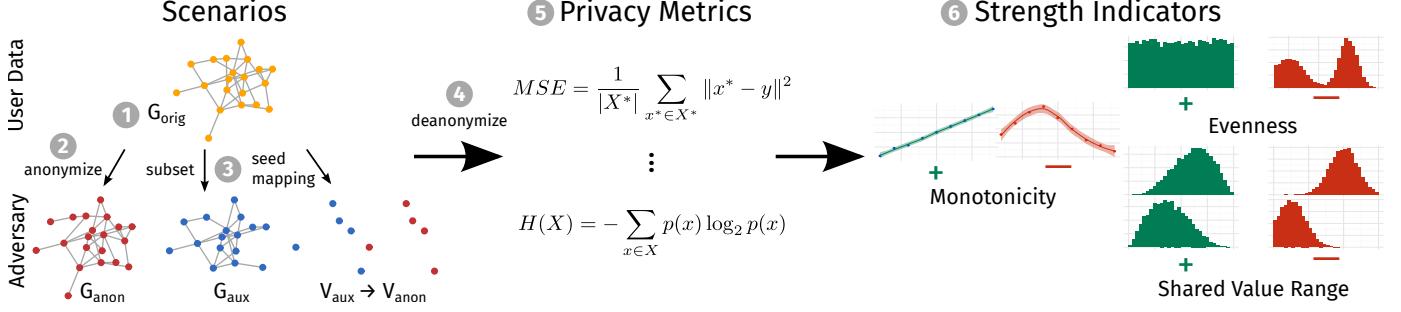


Fig. 1. Our method to evaluate the strength of privacy metrics for graph privacy, adapted from prior work [36], [37].

TABLE 1
Characteristics of the 11 graph datasets used in our experiments, showing their diversity in several key graph statistics.

Dataset	Type	Nodes	Edges	Diameter	Avg. degree	Avg. shortest path	Clustering coefficient	Gini coefficient
dblp	citation	12591	49743	10	8	4.4	0.062	0.66
cora	citation	23166	91500	20	8	5.9	0.117	0.52
arxiv	coauthors	18771	198050	14	21	4.2	0.318	0.61
dnc	communication	2029	5598	8	39	3.4	0.089	0.71
irvine	communication	1899	20296	8	63	3.1	0.057	0.65
manufacturing	communication	167	5784	5	993	1.9	0.541	0.44
caida	computer	26475	53381	17	4	3.9	0.007	0.63
elections	online contact	7118	103675	7	29	3.2	0.125	0.75
pgp	online contact	10680	24316	24	5	7.5	0.378	0.59
google	social	23628	39242	8	3	4.0	0.004	0.66
facebook ego	social	4039	88234	8	2	3.7	0.519	0.54

intact. This method has been shown to be susceptible to de-anonymization attacks [2], [19] and we use it to represent the baseline for non-anonymized graphs.

In addition, we use five anonymization algorithms that change the graph structure: Switch [35], k-Degree Anonymity (k-DA) [15], Differential Privacy (DP) [22], Random Walk (RW) [17], and Bounded t-Means (t-Means) [27]. All algorithms operate on a graph $G_{\text{orig}} = (V_{\text{orig}}, E_{\text{orig}})$, where V_{orig} is the set of nodes (vertices) and E_{orig} is the set of edges. We use $G_{\text{anon}} = (V_{\text{anon}}, E_{\text{anon}})$ to denote the anonymized graph.

Switch [35] randomly chooses two edges in E_{orig} and switches them to change G_{orig} 's structure. This process is repeated $r|E_{\text{orig}}|$ times, where r is the portion of edges to be switched. The k-DA algorithm [15] modifies the graph to achieve k-degree anonymity, so that for each node there are at least $k - 1$ other nodes with the same degree. To achieve this, the algorithm adds edges to E_{orig} until the node degree for all nodes meets the requirement. The DP algorithm [22] transforms the graph to a dK-series and injects noise in it. Then the perturbed dK-series is used to re-construct a differentially private graph. The anonymization level of DP is controlled by the parameter ϵ and the resulting graph satisfies ϵ -differential privacy. The random walk algorithm [17] selects a sequence of vertices starting from v , for each vertex $v \in V_{\text{orig}}$, as a “walking path” of length t . At the end of each sequence, one edge is added between the starting vertex and the ending vertex to change the graph’s structure. The t-Means algorithm [27] first clusters all the vertices in V_{orig} into t clusters based on their degree distances and then matches the degrees of vertices in a cluster to the degree of their center vertex by adding or removing edges.

To apply the anonymization algorithms in our experiments, we use the implementations provided in the open-source software SecGraph [8].

3.3 De-anonymizing Algorithms

Graph de-anonymization attempts to re-identify nodes in an anonymized graph. De-anonymization algorithms commonly assume that the adversary has additional knowledge in the form of an auxiliary graph $G_{\text{aux}} = (V_{\text{aux}}, E_{\text{aux}})$ and a small set of seed mappings between the nodes in V_{aux} and the nodes in V_{anon} .

The auxiliary graph is a sub-graph of the original graph, i.e. $V_{\text{aux}} \subset V_{\text{orig}}$ and $E_{\text{aux}} \subset E_{\text{orig}}$, and the adversary knows the identifiers of nodes in the auxiliary graph. In our experiments, we use the overlap of the auxiliary graph with the original graph as the first way to control the adversary’s strength: a larger overlap results in a stronger adversary.

The set of seed mappings contains some relationships between nodes in the anonymized graph G_{anon} and the auxiliary graph G_{aux} that the adversary is assumed to know. This set of seed mappings is typically small compared to the size of the graph, and most nodes are unmapped. We use the number of seed mappings as the second way to control the adversary’s strength: a larger number of mappings results in a stronger adversary.

The adversary’s goal is to use the auxiliary graph G_{aux} and the seed mappings to generate more mappings between the unmapped nodes in G_{anon} and known nodes in G_{aux} , thus breaching graph privacy. In our experiments, we use six de-anonymization algorithms, all implemented in SecGraph [8]: Narayanan/Shmatikov (NS) [20], Ji/Li/Srivatsa/Beyah (JLSB) [9], Korula/Lattanzi

(KL) [11], Yartseva/Grossglauser (YG) [34], Adaptive De-Anonymization (ADA) [10], and Distance Vector based de-anonymization (DV) [25]. Three algorithms (KL, NS, and YG) rely on seed mappings as a local property of the nodes to be matched, i.e. only inferring among the nodes that are one hop away from seed nodes, while the other three algorithms (ADA, DV, and JLSB) use properties of the global graph structure.

NS is an iterative algorithm. In each step, the adversary randomly chooses an unmapped node v_{anon} . If any of v_{anon} 's neighbors have a known seed mapping to a node v_{aux} , then the set of v_{aux} 's neighbor nodes forms the set of candidate original nodes for v_{anon} . Each candidate is scored according to its node degree, and if the node with the highest score satisfies an eccentricity criterion it is selected as the most likely original node. If the reverse mapping from G_{aux} to G_{anon} of the chosen original node is v_{anon} , the resulting mapping is added to the set of seed mappings and the process is repeated. The inference stops when the set of seed mappings stops growing.

The KL algorithm also bootstraps its inference from seed mappings using a heuristic called similarity witness. The number of similarity witnesses for a pair of known and anonymized nodes is the number of seed mapping relationships between their neighbors. In each step, the pairs with the highest number of similarity witnesses are chosen and added to the set of seed mappings.

Similarly, the YG algorithm randomly selects one unused seed mapping in each iteration and increments the scores of all its neighbor mappings (i.e. mappings between the neighbors of the known nodes and the neighbors of the anonymized node in the selected seed mapping). Once a mapping's score reaches a threshold, it is added to the set of seed mappings.

DV uses a node's distance vector, i.e. the distances between the node and the seed nodes, to describe the structural similarity between anonymized nodes and known nodes. The algorithm maps pairs of nodes that have the highest structural similarity scores.

JLSB uses five heuristics to calculate the structural similarity between anonymized nodes and known nodes: the node degree, neighborhood, top-K reference distance, landmark reference distance, and sampling closeness centrality. These heuristics measure both the local and global properties of each node. The algorithm summarizes these heuristics into a structural similarity score that indicates how much an anonymized node is similar to a known node, thereby finding the most likely mappings.

Similarly, the ADA algorithm calculates structural similarity scores from three centrality measurements: closeness centrality, betweenness centrality, and degree. In addition, ADA takes into account the relative distance similarity (similar to the distance vector in DV) and inheritance similarity (controlling the similarity loss over iterations).

3.4 Metrics for Graph Privacy

Based on our survey of privacy metrics [32], we selected 26 privacy metrics to evaluate in our experiments. These metrics include not only metrics that have already been used in graph privacy, but also metrics from other domains.

Table 2 summarizes the metrics we used. According to the taxonomy in [32], the metrics in our study fall into five categories: uncertainty, information gain/loss, error, similarity, and success.

TABLE 2

Graph privacy metrics used in our experiments. H/L: high (H) or low (L) values indicate high privacy. Per-graph metrics give one privacy value for the entire graph, the other metrics give one privacy value for each node. The *chunk* column indicates whether a metric is negatively affected by SecGraph's chunking (see Section 5.3.2).

Cat.	Metric	H/L	per-graph	gnd.	chunk truth
Uncertainty	Anonymity set size	H	-	-	✓
	Collision entropy	H	-	-	✓
	Conditional entropy	H	-	✓	-
	Conditional privacy	H	-	✓	-
	Entropy	H	-	-	✓
	Inherent privacy	H	-	-	✓
	Max-entropy	H	-	-	✓
	Min-entropy	H	-	-	✓
	Normalized entropy	H	-	-	-
Information gain	Quantiles on entropy	H	-	-	✓
	Amount leaked information	L	✓	✓	-
	Conditional privacy loss	L	✓	✓	✓
	Information surprisal	L	-	✓	-
	Loss of anonymity	L	✓	✓	✓
	Mutual information	L	-	✓	✓
	Pearson correlation	L	-	✓	-
Error	Relative entropy	H	-	✓	-
	Absolute error	H	-	✓	-
	Incorrectness	H	-	✓	-
	Mean squared error	H	-	✓	-
Sim.	% incorrectly classified	H	✓	✓	-
	Normalized variance	H	-	✓	-
Success	Adversary's success rate	L	✓	✓	-
	Adversary's overall success	L	✓	✓	-
	Hiding property	H	✓	-	-
	User-specified innocence	H	✓	-	-

3.4.1 Uncertainty metrics

Uncertainty metrics measure how uncertain the adversary is about her estimate, assuming that higher uncertainty corresponds to better privacy.

For example, the *anonymity set size* for each node indicates how many other nodes the adversary cannot distinguish from this node. We approximate this notion of indistinguishability by counting how many candidate nodes have been assigned a non-zero probability.

All other uncertainty metrics in our study are based on the information theoretic concept of entropy. *Entropy* measures the adversary's uncertainty based on the probabilities assigned to each candidate node and indicates the number of additional bits of information an adversary needs to successfully de-anonymize a node.

Rényi entropy is a generalization of entropy that introduces the parameter α , with the entropy above using $\alpha = 1$. With increasingly larger values of α , the influence of high-probability nodes on the metric increases. For example, *min-entropy* with $\alpha = \infty$ is based only on the node for which the adversary has the highest probability, representing the worst-case privacy. In contrast, *max-entropy* with $\alpha = 0$ is based only on the number of nodes. Because max-entropy does not take into account the adversary's probabilities, it

represents the best-case privacy a user can hope for. *Collision entropy* with $\alpha = 2$ is another variant.

Normalized entropy uses max-entropy to normalize entropy to a common value range of [0,1] that does not depend on the number of nodes. *Quantiles on entropy* aims to mitigate the influence of low-probability outliers on entropy, and thus computes entropy only based on a percentile of the adversary's estimated probabilities. *Conditional entropy* describes the entropy of the true mapping, conditioned on the adversary's estimate.

Inherent privacy, or scaled anonymity set size, is based on entropy and interpreted as the number of additional yes/no questions the adversary has to answer to de-anonymize a node correctly. *Conditional privacy* is similar to inherent privacy, measuring the privacy inherent in the true mapping, given the adversary's estimate.

3.4.2 Information gain/loss metrics

Information gain/loss metrics focus on the amount of information that the adversary gains.

The *amount of information leaked* counts how many nodes the adversary re-identified correctly. *Pearson correlation* computes the correlation between the adversary's estimate and true node mapping. High values indicate a positive correlation and thus mean low privacy.

The remaining metrics in this category are based on information theory. *Information surprisal* focuses on the probability the adversary assigns to the true node and can be interpreted as the amount of surprise felt by the adversary on learning the true mapping.

Mutual information indicates how much information is shared between the adversary's estimate and the true mapping, with more shared information indicating lower privacy. *Loss of anonymity* is the maximum mutual information for any node, and thus indicates the worst-case privacy. *Conditional privacy loss* is a way of normalizing mutual information and can be interpreted as the fraction of privacy lost through the adversary's estimate.

Relative entropy, or Kullback-Leibler divergence, measures the distance between the adversary's estimate and the true mapping, thus indicating how far the adversary's estimate is from the truth.

3.4.3 Error metrics

Error metrics quantify the difference between the adversary's estimate and the true mapping of candidate nodes.

Incorrectness is the expectation of the adversary's estimation error, where successful resp. unsuccessful identification of a node is encoded as 0 resp. 1, and the expectation is computed using the adversary's estimated probabilities.

The *mean squared error* measures the error between the adversary's estimated probability distribution and the true outcome. The *absolute error* measures the difference between the adversary's probability for the true node and the adversary's highest probability, i.e. for node the adversary believes to be the true node.

The *percentage of incorrectly classified* nodes describes the percentage of nodes that the adversary has de-anonymized incorrectly, in relation to the total number of nodes in the graph.

3.4.4 Similarity metrics

Similarity metrics focus on statistical properties that measure similarity between the adversary's estimate and the ground truth.

Normalized variance computes the variance of the difference between the true mapping and the adversary's estimate, normalized by the variance of the true mapping. A higher variance is thought to correlate with higher privacy.

3.4.5 Success metrics

Success metrics evaluate how likely it is that the adversary succeeds. The *adversary's success rate*, or accuracy, indicates the percentage of nodes that the adversary identified correctly. In the variant implemented by SecGraph, this metric indicates the percentage in relation to the number of nodes that the adversary has attempted to de-anonymize, *not* to the total number of nodes in the graph.

The *adversary's overall success rate* indicates the success rate based on the entire graph, not just on the number of de-anonymization attempts.

The *hiding property* counts the number of nodes for which the adversary's largest probability is below a specified threshold. This metric indicates the number of nodes that the anonymizing algorithm has successfully protected from the adversary. *User-specified innocence* counts the number of nodes for which the adversary's estimated probability for the true outcome is below a specified threshold, indicating the number of nodes that can reasonably claim that the adversary's de-anonymization attempt was not reliable.

3.5 Computation of Monotonicity Scores

To evaluate the strength of privacy metrics, we require that they are monotonic, i.e. that they indicate higher privacy levels for stronger adversaries. For example, we expect that, with increasing adversary strength, the values of the adversary's success rate (as a lower-better metric) decrease, and that the values of entropy (as a higher-better metric) increase.

We use the algorithm originally proposed in [30] to compute monotonicity scores. The algorithm uses two statistical tests (t-test and rank-sum test) to compare the mean metric values for each pair of adjacent adversary strengths. If the difference between the means is statistically significant and indicates a change in the expected direction, the algorithm increases the metrics' monotonicity score by 1. If the difference indicates a change in the wrong direction, the algorithm subtracts 1 from the monotonicity score. If the changes in metric values change direction, e.g. increasing for one pair and decreasing for the next, the algorithm reduces the score by 2 because such a peak may indicate the same privacy levels for both strong and weak adversaries and is thus undesirable. A metric's final monotonicity score is the average of the scores for the two statistical tests, normalized to [0, 1].

3.6 Computation of Additional Criteria

In addition to monotonicity, two additional criteria are useful to judge the strength of privacy metrics, especially if they will be used for within-scenario comparisons and between-scenario comparisons [36].

Within-scenario comparisons compare the privacy levels of different nodes within the same graph. Ideally, the values of metrics for within-scenario comparisons should be spread evenly over the entire value range (*evenness*). We compute the evenness of a metric’s value range based on the Cramér-von Mises criterion, which measures the goodness of fit between the uniform distribution $U(0,1)$ and the normalized metric values for all adversary strengths. We normalize the Cramér-von Mises criterion by the number of metric values to offset the influence of the number of samples.

Between-scenario comparisons compare privacy levels between different graphs, and ideal metrics use the same value range regardless of the graph’s characteristics (*shared value range*). We formalize this notion by computing the portion of the global value range for each metric that is used when the metric is applied to a specific graph.

4 EXPERIMENTS

We conducted extensive experiments to evaluate monotonicity, evenness, and shared value range in a wide range of scenarios. In this section, we give details on the implementation and availability of software, the parameter settings we used, and how we controlled the statistical significance of our results.

4.1 Implementation

We have implemented our framework for the evaluation of privacy metrics in Python. Our implementation of privacy metrics follows the description in [32] and relies on Python libraries numpy, scipy, and scikit-learn.

For anonymization and de-anonymization algorithms, we used the open-source software SecGraph [8]. However, SecGraph’s implementation of de-anonymization algorithms only outputs the adversary’s node mapping and success rate. To be able to compute other privacy metrics, we modified the de-anonymization methods to additionally output the ground truth and the adversary’s estimated probability distribution. The probability distribution is based on the adversary’s scores for candidate nodes. When the adversary tries to map a node v_{anon} to a candidate in a set of candidate nodes $\{v_{\text{aux}}^1, v_{\text{aux}}^2, \dots\}$, the random variable X describes the probabilities for each candidate in the set. The probability for each candidate node is defined as its score divided by the sum of all scores in the set. In this way, the adversary creates one probability distribution for each node she attempts to re-identify.

To automate our experiments and make use of computing resources available to us, we packaged our framework in a Docker container and used Boinc and boinc2docker to distribute the computation to around 60 PCs available in student labs. Our source code is available at CodeOcean [33].

4.2 Parameter Settings

We reproduce parameter settings for the anonymization and de-anonymization algorithms as much as possible from [8]. However, the parameters given in the SecGraph paper do not always match the implementation in their software, and the software parameter choices are not documented. In these

TABLE 3
Parameter settings used in our experiments.

Anonymizers	
DP	$\epsilon = 1$
k-DA	$k = 5$
RW	distance = 2
Switch	fraction $r = 0.05$
t-means	max size = 30
De-anonymizers	
ADA	$\theta = 0$, chunk size = 100, $\epsilon = 0.5$, weights: $w_{\text{distance}} = 0.6$, $w_{\text{structural}} = 0.2$, $w_{\text{inheritance}} = 0.2$
DV	$\theta = 0$, chunk size = 100
JLSB	$\theta = 0$, chunk size = 100, weights: $w_{\text{degree}} = 0.3$, $w_{\text{neighbor}} = 0.3$, $w_{\text{ref distance}} = 0.4$
KL	$\theta = 1$, chunk size = 100
NS	$\theta = 0.5$
YG	$\theta = 2$
Seed numbers	5, 10, 20, 35, 50, 100 (default 50)
Auxiliary ratios	0.6, 0.7, 0.8, 0.85, 0.9, 0.95 (default 0.85)

cases, we chose parameter settings that seemed reasonable to us. Table 3 summarizes our parameter settings. The last two rows show the sequence of parameters we used to increase the adversary’s strength. We varied seed numbers and auxiliary ratios in independent sets of experiments and used the value given as default for the variable that was kept constant.

Our experiments are based on random inputs, for example the selection of seed mappings and the overlap of the auxiliary graph. To obtain statistically significant results, we replicated the experiments until the relative error for the metric values for each combination of dataset, anonymizer, and deanonymizer was below 5%. 100 replications were sufficient in most cases, but in some cases we computed up to 50.000 replications.

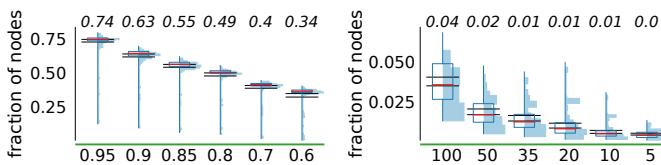
5 RESULTS ON THE MONOTONICITY OF GRAPH PRIVACY METRICS

Our experiments have yielded results for 792 individual scenarios, i.e. 792 combinations of dataset, anonymizer, de-anonymizer, and adversary strength type, with each scenario composed of results for six adversary strength levels. In this section, we discuss our results for monotonicity and analyze the factors that influence monotonicity. Results for the additional criteria evenness and shared value range are in the next section.

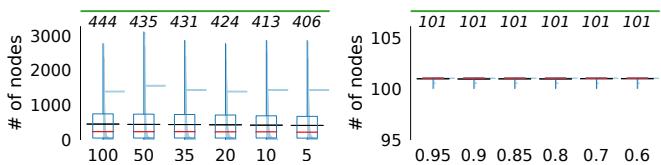
5.1 Overview of Results

Figure 2 shows detailed results for three metrics in six of our 792 scenarios. Each subfigure shows a sequence of box plots, one for each adversary strength level, with rotated histograms indicating the distribution of metric values collected from all replications. Black horizontal lines indicate confidence intervals for the mean, italic values on top of each box indicate the mean value, and the green line at the top (resp. bottom) indicates whether higher (resp. lower) values indicate higher privacy.

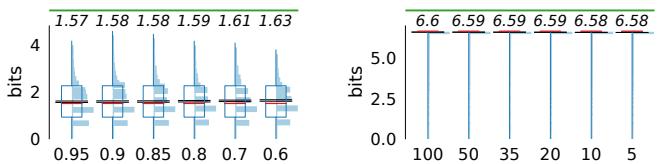
According to our monotonicity criterion, we expect that boxes on the right-hand side of the plots are closer to the green line than boxes on the left, i.e. that metrics indicate higher privacy for lower adversary strengths. The



(a) *Adversary's success rate*, DV, ID removal, Irvine, varying aux ratio. Monotonicity=1.0 (b) *Adversary's success rate*, KL, t-means, Caida, varying seed numbers. Monotonicity=1.0



(c) *Anonymity set size*, NS, Switch, Elections, varying seed numbers. DBLP, varying aux ratio. Monotonicity=0.14 (d) *Anonymity set size*, JLSB, RW, DBLP, varying aux ratio. Monotonicity=0.41



(e) *Entropy*, YG, DP, Arxiv, varying aux ratio. Monotonicity=0.46 (f) *Entropy*, ADA, k-DA, DNC, varying seed numbers. Monotonicity=0.11

Fig. 2. Detailed results for a selection of metrics.

first row shows the *adversary's success rate* in two scenarios (Figures 2a–2b). In both cases, the metric indicates a clear decrease in the success rate from left to right, with a monotonicity score of 1.0.

The plots in the second row show the *anonymity set size* in two scenarios (Figures 2c–2d). In Figure 2c, the low monotonicity score results from a change in the wrong direction: instead of the *anonymity set size* getting larger with decreasing adversary strength, it is actually getting smaller.

The lack of variation in Figure 2d is due to an undocumented implementation detail in SecGraph: for all de-anonymizers that use the global graph structure (ADA, DV, JLSB), the SecGraph implementations process the graph in chunks to keep the runtime within reasonable limits (as the plot shows, we chose a chunk size of 100). This chunking does not have a large effect on the *adversary’s success rate* because the nodes in the anonymized and auxiliary graphs are ordered by their degrees before chunking. However, metrics that use the adversary’s probabilities are skewed by this artificial limit.

The third row (Figures 2e–2f) shows entropy in two scenarios. *Entropy* measures the adversary's uncertainty, and we expect that *entropy* increases with higher anonymization levels. In Figure 2e, the general trend of the metric is in the right direction. However, the monotonicity score is only a medium 0.46 because some of the adversary strength levels have no statistically significant difference. Figure 2f shows how *entropy* is affected by SecGraph's chunking strategy: the artificial limit of 100 candidates for each node results in almost constant values for *entropy* throughout.

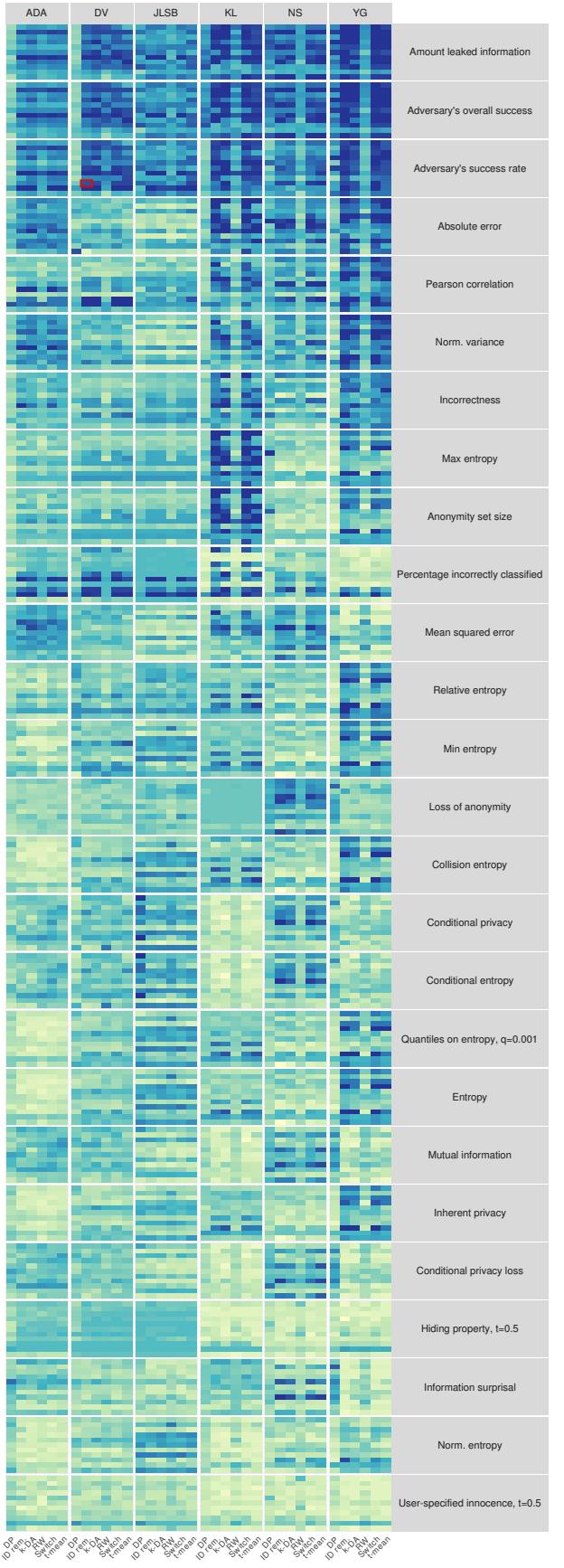


Fig. 3. Heat map visualizing monotonicity scores for all privacy metrics in our study, depending on the deanonymization algorithm (top), anonymization algorithm (bottom), and graph dataset (left). Each cell averages results for varying seed numbers, varying aux ratios, and all replications. Light yellow colors indicate low monotonicity (weak metric), and dark blue colors indicate high monotonicity (strong metric).

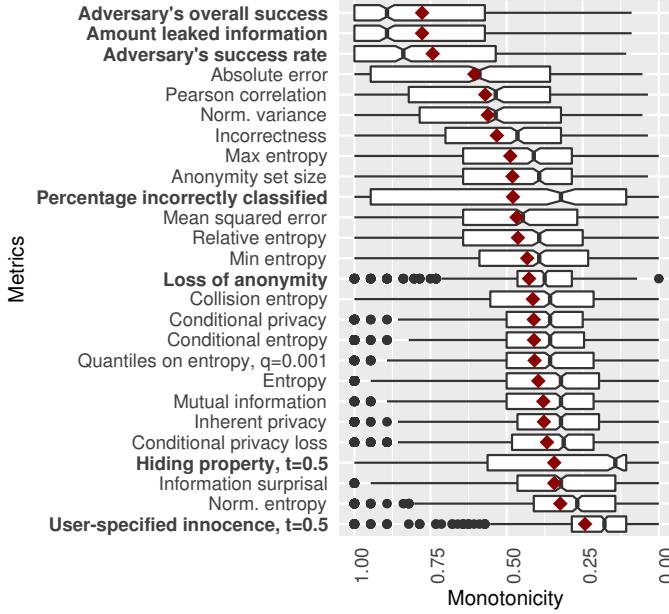


Fig. 4. Box plot showing the ranking of privacy metrics according to their monotonicity scores. Each box summarizes monotonicity for all anonymization and deanonymization algorithms and all graph datasets. Per-graph metrics are marked in bold.

We have applied the algorithm described in Section 3.5 to summarize the individual results from Figure 2 into monotonicity scores. Figure 3 shows these monotonicity scores on a heat map. Each field in the heat map represents one set of box plots from Figure 2. As an example, we have highlighted the field corresponding to Figure 2a with a red outline. The heat map shows that only few metrics have high monotonicity, most notably the *adversary's success rate* and the *amount of leaked information*. These metrics are based on the count of nodes that the adversary has successfully re-identified.

Most other metrics have medium or low monotonicity. This includes almost all metrics that are based on the adversary's probability distribution, for example *entropy* and the metrics derived from *entropy*.

Figure 4 shows the monotonicity scores for all metrics in a box plot where each box combines the monotonicity scores for all anonymization algorithms, all deanonymization algorithms, and all graph datasets. The plot is ordered according to the average score. We can see that most metrics have monotonicity scores below 0.5, which indicates that they are not monotonic for most anonymization levels and thus should be used with caution, if at all. In addition, many metrics that perform well in other domains do not perform well when applied to graph privacy, for example *normalized entropy* and the *anonymity set size*. We investigate the reasons for this in the following two sections.

5.2 Analysis of Factors Influencing Monotonicity

We have defined monotonicity as a property of privacy metrics, so we expect that the metrics should be the main determining factor for the value of monotonicity. However, we have evaluated monotonicity in complex scenarios that include different graph datasets, anonymization algorithms,

TABLE 4
MARS model showing which factors influence monotonicity, in order of importance

monotonicity =		
0.26	*	metric Relative entropy
+	0.095	* metric Percentage incorrectly classified
+	0.11	* metric Pearson correlation
+	0.2	* metric Norm. variance
+	0.19	* metric Min-entropy
+	0.064	* metric Mean squared error
+	0.1	* metric Max-entropy
+	0.12	* metric Loss of anonymity
+	0.058	* metric Incorrectness
+	0.16	* metric Anonymity set size
+	0.11	* metric Amount leaked information
+	0.41	* metric Adversary's success rate
+	0.37	* metric Adversary's overall success
+	0.41	* metric Absolute error
+	0.24	* anonymizer IDrem.
+	0.089	* anonymizer k-DA
+	0.082	* anonymizer Switch
+	0.09	* anonymizer t-mean
+	0.082	* de-anonymizer JLSB
+	0.026	* de-anonymizer YG
+	0.037	* de-anonymizer YG
-	0.066	* max(0, dataset pgp - 0)
+	5.4e-06	* max(0, 10680 - nodes)
-	5.7e-06	* max(0, nodes - 10680)
-	2.9e-06	* max(0, 13838 - edges)
+	3.6e-07	* max(0, edges - 13838)
+	8.3e-11	* max(0, 7.3e+08 - claws)
+	1.7e-11	* max(0, claws - 7.3e+08)
-	0.0083	* max(0, 7.7 - avg. degree)

and deanonymization algorithms. Each of these factors may influence the value of monotonicity as well. To find out to what extent the variation in monotonicity can be explained by each of these factors, we used multivariate adaptive regression splines (MARS) [5] to analyze how much each of them contributes to the final monotonicity score. MARS is more suitable than linear regression in this case because it accounts for non-linearity and interactions between the variables and automatically selects the most important variables to include in the model.

Table 4 shows the contribution of each factor according to the MARS analysis. The results show that metrics are indeed the most important contributors and have by far the largest influence on the monotonicity score. In contrast, the influence from anonymizing/de-anonymizing algorithms and graph statistics is much smaller. In many cases, the influence was so small that the variables were not important enough to be included in the model at all.

We fit a linear regression to the data to confirm this finding and found that all graph statistics combined account for less than 2% of the variation in monotonicity ($R^2 = 0.013$), and the anonymizer, de-anonymizer, dataset and graph type combined account for less than 8% ($R^2 = 0.074$).

We conclude that despite our complex evaluation setup, monotonicity is primarily influenced by the choice of metric, and not by the dataset or anonymization/de-anonymization algorithms. Because we have evaluated a wide range of datasets, anonymizers, and de-anonymizers, we are confident that our results on the monotonicity of graph privacy metrics hold also for other datasets and algorithms.

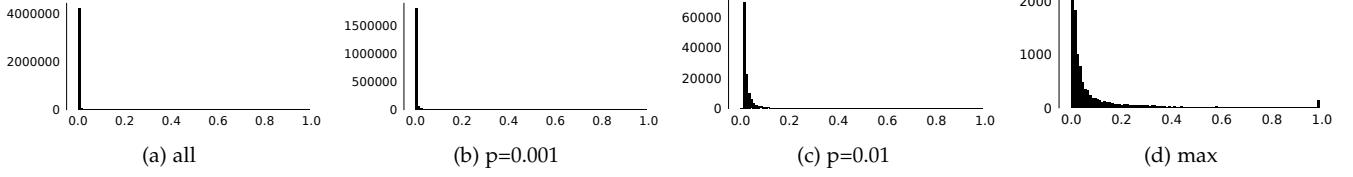


Fig. 5. Distribution of the adversary's estimate for elections, NS, k-DA. Figure (a) shows all values, (b) and (c) cut off all probability values smaller than p , and (d) shows the probability distribution for the most likely candidate node.

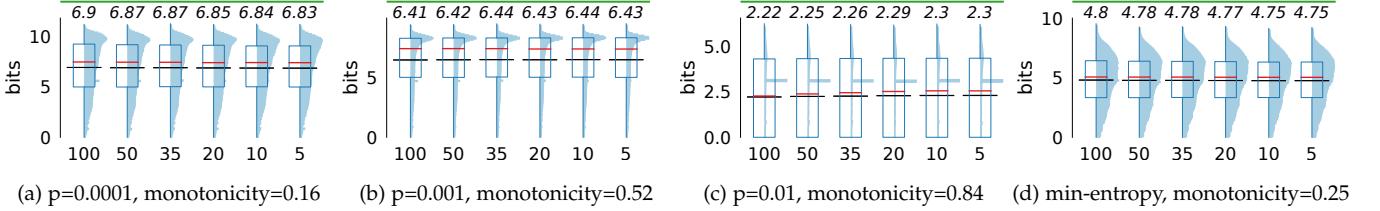


Fig. 6. *Entropy* based on the adversary's estimated probabilities without probabilities below p , for elections, NS, k-DA.

5.3 Analysis of Non-monotonic Metrics

Our results show that some metrics that are popular and monotonic in other application domains are not monotonic when measuring graph privacy. In this section we analyze why this is the case for information theoretic metrics based on *entropy* and the *anonymity set size*. We also analyze the behavior of the *adversary's success rate* in detail.

5.3.1 Entropy affected by low-probability candidates

One reason why *entropy* and metrics derived from entropy have low monotonicity scores may be that low-probability candidates have a large influence on *entropy* [4], [18]. To find out whether this is the case here, we plot the adversary's probability distribution using a histogram with 100 bins (Figure 5a). The histogram clearly shows that most values are in the left-most bin, indicating that most probabilities are below 1%, which confirms the presence of a large number of low-probability candidates.

As suggested in [4], a possible way of dealing with these low-probability candidates is to calculate *entropy* based on a quantile of the adversary's probability distribution, i.e. to remove a certain portion of low-probability candidates before calculating the entropy value. To evaluate the effect of this, we define versions of *entropy* that cut off probability values below p before the calculation. Figure 6 shows the detailed results and monotonicity scores for three of these modified entropies, with $p = \{0.0001, 0.001, 0.01\}$, and Figures 5b–5c show the corresponding histograms for the adversary's probability distribution. We can see that removing probability values below 0.0001 does not improve the monotonicity score for *entropy*, 0.001 shows a moderate improvement, and 0.01 shows a clear improvement.

Depending on the anonymization algorithm, removing small probability values corresponds to removing a certain percentile of the adversary's probability distribution: removing probability values below 0.0001 corresponds to removing 0.16% of the probability distribution, removing $p < 0.001$ corresponds to removing 55%, and removing $p < 0.01$ corresponds to 97% of the probability distribution.

Even though *entropy* in the last case is monotonic, it is questionable whether it still makes sense as a privacy metric because it only evaluates the adversary's uncertainty based on the top 3% of her probability distribution.

Another way of dealing with low-probability candidates is to use *min-entropy*, a variant of *entropy* that is calculated based only on the most likely candidate. However, as Figure 6d shows, *min-entropy* shows the same non-monotonic behavior as *entropy*. To explain why, we plot the distribution of the adversary's probability for the most likely candidate for each node in Figure 5d. The figure shows that for the majority of nodes, the adversary makes a mapping decision based on a probability smaller than 1% (the left-most bin). Just like *entropy*, *min-entropy* is therefore influenced by low-probability candidates.

As a result, entropy-based metrics do not seem to be suitable to evaluate graph privacy, even if only a certain percentile of the adversary's distribution is used in their calculation.

5.3.2 Metrics affected by SecGraph's chunks

We have shown in Figure 2d that the anonymity set size can be influenced by SecGraph's strategy of processing graphs in chunks. This chunking does not necessarily result in low monotonicity scores because the metric values are often indistinguishable from one adversary strength to the next, which results in a medium monotonicity. However, the semantics and interpretation of the metric are affected: in the case of the *anonymity set size*, the interpretation is no longer "all nodes that the adversary cannot distinguish," but instead "all nodes that the adversary cannot distinguish and that are in a group of 100 nodes with similar node degrees." This change in interpretation can make the metric unsuitable for evaluating graph privacy.

To analyze whether the *anonymity set size* is the only metric affected by SecGraph's chunking, we plot the maximum values for each metric separately for each de-anonymizing algorithm. If a metric is affected by chunking, the four deanonymizers that use chunking (ADA, DV, JLSB, KL)

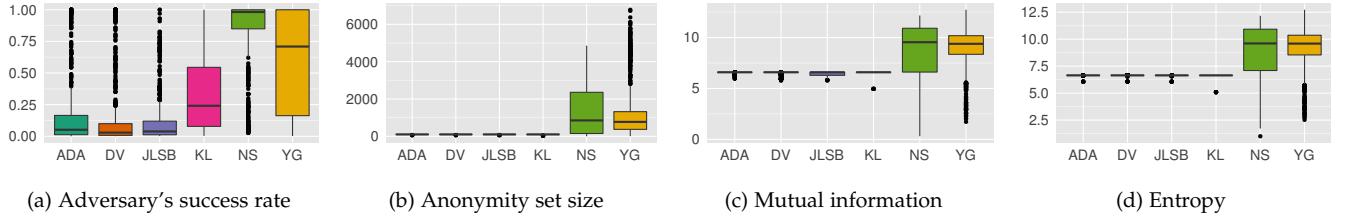


Fig. 7. Maximum values for four metrics across all datasets and anonymizers, by de-anonymizing algorithm. The *adversary's success rate* is not affected by SecGraph's chunking, whereas the other three metrics are: their maximum values have an artificial upper limit.

should show a consistent upper limit, whereas NS and YG should not. Figure 7 shows these plots for four example metrics. It is clear that the *anonymity set size*, *mutual information*, and *entropy* are affected by chunking, but the *adversary's success rate* is not. Table 2 (column *chunk*) indicates which metrics are negatively affected by SecGraph's chunking.

5.3.3 Adversary's Success Rate vs Overall Success Rate

We have already mentioned the difference between the *adversary's success rate* and *adversary's overall success* in Section 3.4: success rate is based on the number of attempted de-anonymizations, whereas overall success rate is based on the total number of nodes in the graph. We will show in this section why this difference can lead to misjudgment of privacy levels when comparing different de-anonymization algorithms.

We have studied two groups of de-anonymization algorithms: global algorithms (ADA, DV, JLSB) re-identify nodes based on the global graph structure, whereas local algorithms (KL, NS, YG) use only local information such as the number of neighbors. Local algorithms generally bootstrap from the given seed mappings and iteratively add nodes to the set of seeds, stopping when no more nodes can be added. As a result, local algorithms attempt to re-identify fewer nodes than global algorithms, and this increases their success rate (which is based on attempted nodes) compared to global algorithms.

Figure 8 compares the two metrics for each of the six de-anonymization algorithms. We can see that the *adversary's overall success* (light color) is nearly equal for all six algorithms, especially when comparing the median value indicated by the line in each box. For the global algorithms, the *adversary's success rate* is very close to the *adversary's overall success*. In contrast, the local algorithms report a much higher success rate than overall success rate.

There is no doubt that the *adversary's success rate* can be a useful metric, for example to highlight the potential of a deanonymizer if the right kind of auxiliary information is available to the adversary. However, it is unsuitable for comparing different deanonymizing algorithms because it reports inflated success rates for local algorithms.

6 RESULTS ON ADDITIONAL CRITERIA

Our analysis so far has focused on the monotonicity of privacy metrics. However, as the comparison between the *adversary's success rate* and *overall success rate* has shown, monotonicity is not always a sufficient criterion to select privacy metrics. Prior work has identified evenness and

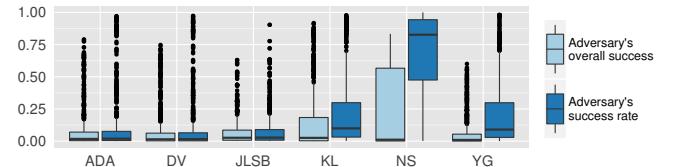
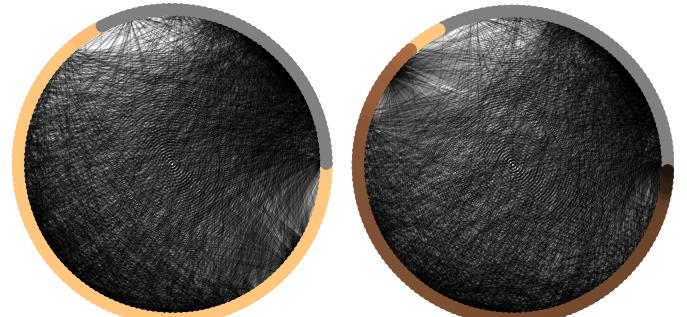


Fig. 8. Comparison between the *adversary's success rate* and the *adversary's overall success*. De-anonymizers that use local information (KL, NS, YG) have an inflated success rate because it is normalized with the number of attempted re-identifications.



(a) *Incorrectness*, Manufacturing-DV, DP. Evenness=0.14 (b) *Pearson correlation*, Manufacturing-DV, DP. Evenness=0.89

Fig. 9. Metric values visualized on the edges of a circular graph layout. Light colors: high privacy, gray: de-anonymization not attempted.

shared value range as additional criteria for metric selection especially for within-scenario comparisons (evenness) and between-scenario comparisons (shared value range) [36].

6.1 Evenness

To illustrate the requirement for evenness, Figure 9 compares two privacy metrics applied to the same scenario. Each subfigure shows the graph dataset in a circular layout, with the nodes colored according to their privacy. Light colors indicate high privacy, and nodes for which deanonymization has not been attempted are colored gray. *Incorrectness* on the left (Figure 9a) has a low evenness score (0.14), which can be seen by the absence of medium and dark colors. In contrast, *pearson correlation* on the right (Figure 9b) has a high evenness score (0.89), which is visible in the clear representation of light, medium, and dark colors. Metrics with a high evenness score thus allow to analyze which nodes in a graph have better privacy protection than others.

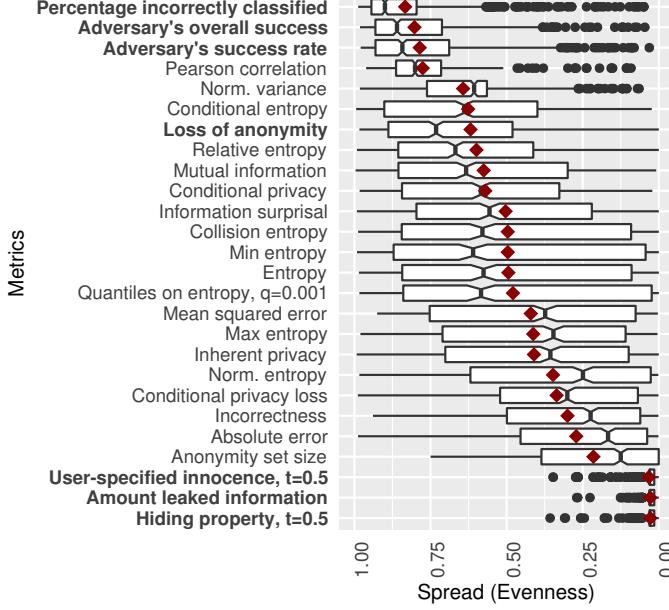


Fig. 10. Box plot showing the ranking of privacy metrics according to their evenness scores.

We have computed evenness scores according to the description in Section 3.6 for all 792 scenarios. Figure 10 shows a ranking of privacy metrics according to their average evenness score, with the boxes indicating the distribution of scores across all scenarios. Among the metrics with an average evenness score above 0.5, four metrics also have a monotonicity score above 0.5: *adversary’s overall success*, *adversary’s success rate*, *pearson correlation*, and *normalized variance*.

6.2 Shared Value Range

To illustrate the requirement for a shared value range, we show two metrics, *pearson correlation* and *anonymity set size*, in four scenarios each in Figure 11. The top row shows *pearson correlation* with a consistent value range of [0,1] in every scenario. In contrast, the bottom row shows the *anonymity set size*, where each scenario uses a different portion of the global value range, depending on the size of the graph, the anonymization/de-anonymization algorithms and how they are implemented. This indicates that the *anonymity set size* is less suitable for comparisons between scenarios than *pearson correlation*.

Figure 12 shows the distribution and average values for the shared value range score across all scenarios. Only three metrics have a shared value range score above 0.5 as well as a monotonicity score above 0.5: *pearson correlation*, *absolute error*, and *normalized variance*.

7 RECOMMENDATIONS: METRICS SUITES

Based on our experimental results and the analysis in the previous sections, we can see that no single metric is the ideal metric for all scenarios. The metrics with the highest monotonicity are per-graph metrics, so they do not provide information about the privacy levels of individual nodes in a graph. Most of the other metrics have very low monotonicity, and only seven metrics have an average monotonicity

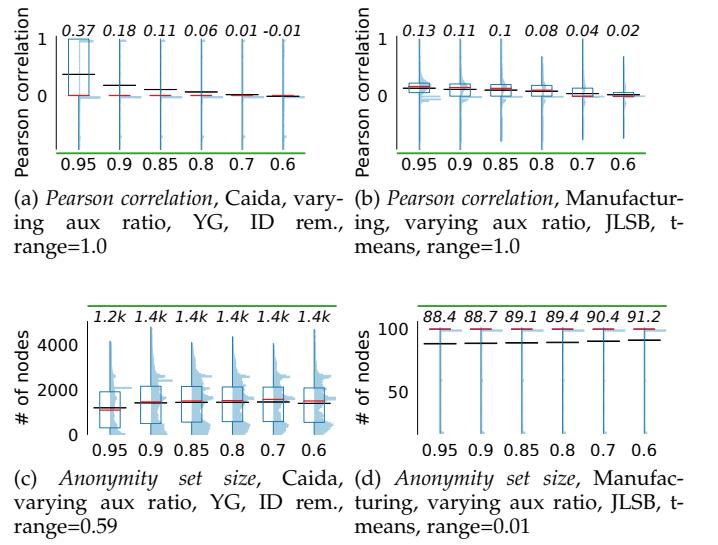


Fig. 11. Shared value range: *pearson correlation* in the top row has a consistent value range across scenarios, while the *anonymity set size* in the bottom row uses a different value range (y axis) for each scenario.

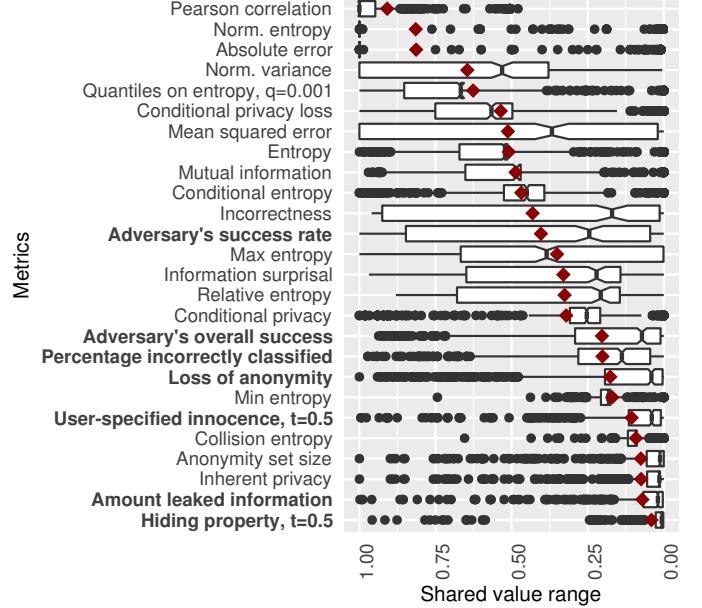


Fig. 12. Box plot showing the ranking of privacy metrics according to their shared value range scores.

score of 0.5 or above. When considering evenness and shared value range in addition to monotonicity, only two metrics have high scores in all three criteria (*pearson correlation* and *normalized variance*). Finally, another important consideration is the semantics of metrics and their interpretation. In this regard, it is desirable to measure different aspects of privacy, for example in terms of the categories described in Section 3.4.

Therefore, our recommendation is to combine several privacy metrics in a metrics suite. To allow for easy comparisons between anonymizing or de-anonymizing algorithms, and to be able to decide which provides the best or worst privacy levels, it is desirable to combine the metrics in a suite into a single number. This is not a straightforward task

because metrics have different scales of measurement and different directions (for some, higher values indicate higher privacy, for others, lower values). A simple average is therefore unlikely to yield good results, and any normalization has to carefully consider direction and value range.

7.1 Combining Metrics in a Metrics Suite

The problem of combining metrics is similar to a problem in operations research: given a set of alternatives and a number of criteria for each alternative, which alternative is best? Many methods have been proposed for multi-criteria decision analysis [29]. The simplest method is the weighted sum model (WSM), which ranks alternatives according to a weighted sum of the criteria values. However, this method is only applicable if the criteria are measured on the same scale and in the same units. The weighted product model (WPM) instead computes the relative importance Q_i of each alternative i using a weighted product. Applied to the case of privacy metrics, each criterion j corresponds to one privacy metric, and the set of alternatives is the set of scenarios (e.g., anonymization algorithms or parameter settings) that need to be compared in terms of their privacy. According to [29], the relative importance of each alternative i is computed as

$$Q_i = \prod_{j=1}^n (\bar{x}_{ij})^{w_j},$$

where w_j is the weight assigned to each metric, and \bar{x}_{ij} is a normalization of the metric value x_{ij} such that

$$\bar{x}_{ij} = \frac{x_{ij}}{\max_i x_{ij}}$$

if higher values indicate higher privacy, and

$$\bar{x}_{ij} = \frac{\min_i x_{ij}}{x_{ij}}$$

if lower values indicate higher privacy. The advantages of WPM are that different units of measurement are effectively eliminated due to the use of multiplication instead of addition, that the built-in normalization can integrate higher-better and lower-better metrics, and that it does not suffer from rank reversals that can occur with additive methods [28].

7.2 Choosing Metrics for a Metrics Suite

As we have already discussed above, every metric in a metrics suite should have a monotonicity score of at least 0.5. The choice of metrics can then depend on the specific demands of the application and might include metrics with high evenness, metrics with high shared value range, metrics from different categories, and metrics that are easy to interpret in the application context.

In our experiments, the seven metrics that have high monotonicity scores are the *adversary's overall success*, *amount leaked information*, *adversary's success rate*, *absolute error*, *pearson correlation*, *normalized variance*, and *incorrectness*. This list includes metrics with high evenness and shared value range, and cover fours different categories (information gain/loss, error, similarity, success probability). The list also includes the *adversary's success rate* which is the most common metric to evaluate graph privacy.

7.3 Evaluating the Performance of Metrics Suites

To find out which combination of metrics results in the best metrics suite, we have evaluated all metrics suites resulting from combinations of the top-7 monotonic metrics. A good metrics suite should create a monotonic ranking of alternatives. Based on our experimental data, we can evaluate the monotonicity of metrics suites as follows. Each of our 792 scenarios consists of six adversary strength levels, with increased strength defined either by an increase in the number of seed mappings or by an increase in the overlap of the auxiliary graph. We can therefore use each scenario as one set of six alternatives, using the increasing strength levels as the ground truth for how the alternatives should be ranked. For each set of alternatives, we used the mean metric values across all replications as x_{ij} and then count how many of the six strength levels have been ranked monotonically by the metrics suite score Q_i .

We find that most metrics suites create monotonic rankings for more than 80% of scenarios, and only metrics suites that include none of the top-3 metrics create monotonic rankings for less than 70% of scenarios. In comparison, the best individual metrics (*adversary's overall success* and *amount information leaked*) create monotonic rankings for 88.2% of scenarios.

Table 5 summarizes the composition and weights for four of the top-scoring metrics suites and the best individual metric. When choosing all weights w_j to be equal, the best metrics suite, consisting of *pearson correlation*, *adversary's overall success*, *normalized variance* and *amount leaked information*, ranks 88.6% of scenarios monotonically. Importantly, this is better than the result for the best individual metric.

When choosing unequal weights, the best metrics suite we have been able to construct consists of *pearson correlation*, *normalized variance*, *incorrectness*, *amount leaked information*, *adversary's success rate*, *adversary's overall success*, and *absolute error* with weights $w_i = 0.1, 0.1, 0.1, 0.25, 0.1, 0.25, 0.1$. This metrics suite creates monotonic rankings for 89% of scenarios.

In summary, we find that combining privacy metrics in a metrics suite can improve monotonicity above the monotonicity of individual metrics. This is an important result that shows a concrete method towards more monotonic and thus more accurate evaluations of privacy.

8 CONCLUSION

In this paper, we analyzed the strength of 26 privacy metrics for graph privacy in terms of their monotonicity, evenness, and shared value range. We conducted extensive experiments on 11 public graph datasets, using 6 anonymization algorithms and 6 de-anonymization algorithms. We found that most metrics are not monotonic when applied in graph privacy, including several metrics that are popular and monotonic in other fields. Our detailed analysis of the strengths and weaknesses of these metrics led us to propose metrics suites, that is, combinations of privacy metrics that can combine the strengths and mitigate the weaknesses of individual metrics. To the best of our knowledge, we were the first to apply techniques from multi-criteria decision analysis to privacy measurement and found that the resulting metrics suites can indeed increase monotonicity above

TABLE 5
Composition of four metrics suites (S2 to S5) and the percentage of monotonic rankings resulting from their aggregation with WPM, compared with the best individual metric (S1).

Metrics	Weights	% mono
S1 Adversary's overall success	equal	88.2
S2 Adversary's overall success, Norm. variance	equal	88.6
S3 Pearson correlation, Adversary's overall success, Norm. variance, Amount leaked information	equal	88.6
S4 Pearson correlation, Adversary's overall success, Norm. variance, Amount leaked information, Incorrectness	0.1, 0.35, 0.1, 0.35, 0.1	88.7
S5 Pearson correlation, Norm. variance, Incorrectness, Amount leaked information, Adversary's success rate, Adversary's overall success, Absolute error	0.1, 0.1, 0.1, 0.25, 0.1, 0.25, 0.1	89.0

the monotonicity of the best individual metric. This result opens up a new line of research that may lead to significant improvements in privacy measurement.

ACKNOWLEDGMENTS

This work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) grant EP/P006752/1 and used the ARCHER UK National Supercomputing Service (<http://www.archer.ac.uk>).

REFERENCES

- [1] J. Alexander and J. Smith, "Engineering Privacy in Public: Confounding Face Recognition," in *3rd International Workshop on Privacy Enhancing Technologies (PET)*. Dresden, Germany: Springer LNCS, volume 2760, Mar. 2003, pp. 88–106.
- [2] L. Backstrom, C. Dwork, and J. Kleinberg, "Wherefore Art Thou R3579x?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography," in *Proceedings of the 16th International Conference on World Wide Web*, ser. WWW '07. Banff, Alberta, Canada: ACM, May 2007, pp. 181–190.
- [3] E. Bertino, D. Lin, and W. Jiang, "A Survey of Quantification of Privacy Preserving Data Mining Algorithms," in *Privacy-Preserving Data Mining: Models and Algorithms*. Springer Advances in Database Systems, Jul. 2008, no. 34, ch. 8, pp. 183–205.
- [4] S. Clauf and S. Schiffner, "Structuring Anonymity Metrics," in *Proceedings of the 2nd ACM Workshop on Digital Identity Management (DIM)*, Alexandria, VA, USA, Nov. 2006, pp. 55–62.
- [5] J. H. Friedman, "Multivariate Adaptive Regression Splines," *The Annals of Statistics*, vol. 19, no. 1, pp. 1–67, 1991.
- [6] S. Ji, W. Li, S. Yang, P. Mittal, and R. Beyah, "On the Relative De-Anonymizability of Graph Data: Quantification and Evaluation," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, San Francisco, CA, USA, Apr. 2016, pp. 1–9.
- [7] S. Ji, P. Mittal, and R. Beyah, "Graph Data Anonymization, De-Anonymization Attacks, and De-Anonymizability Quantification: A Survey," *IEEE Communications Surveys Tutorials*, vol. 19, no. 2, pp. 1305–1326, Secondquarter 2017.
- [8] S. Ji, W. Li, P. Mittal, X. Hu, and R. Beyah, "SecGraph: A Uniform and Open-source Evaluation System for Graph Data Anonymization and De-anonymization," in *Proceedings of the 24th USENIX Security Symposium (USENIX Security)*, Washington, D.C., Aug. 2015, pp. 303–318.
- [9] S. Ji, W. Li, M. Srivatsa, and R. Beyah, "Structural Data De-anonymization: Quantification, Practice, and Implications," in *Proceedings of the 2014 ACM Conference on Computer and Communications Security*, Scottsdale, AZ, USA, Nov. 2014, pp. 1040–1053.
- [10] S. Ji, W. Li, M. Srivatsa, J. S. He, and R. Beyah, "Structure Based Data De-Anonymization of Social Networks and Mobility Traces," in *Proceedings of the 17th International Conference on Information Security*, Hong Kong, China, Oct. 2014, pp. 237–254.
- [11] N. Korula and S. Lattanzi, "An Efficient Reconciliation Algorithm for Social Networks," *Proceedings of the VLDB Endowment*, vol. 7, no. 5, pp. 377–388, Jan. 2014.
- [12] R. Kumar, J. Novak, and A. Tomkins, "Structure and Evolution of Online Social Networks," in *Link Mining: Models, Algorithms, and Applications*, 2010, pp. 337–357.
- [13] J. Kunegis, "KONECT: The Koblenz Network Collection," in *Proceedings of the 22nd International Conference on World Wide Web*. Rio de Janeiro, Brazil: ACM, May 2013, pp. 1343–1350.
- [14] J. Leskovec and R. Sosić, "SNAP: A General-Purpose Network Analysis and Graph-Mining Library," *ACM Transactions on Intelligent Systems and Technology*, vol. 8, no. 1, pp. 1:1–1:20, Jul. 2016.
- [15] K. Liu and E. Terzi, "Towards Identity Anonymization on Graphs," in *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD)*, Vancouver, Canada, Jun. 2008, pp. 93–106.
- [16] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, "Recommender Systems with Social Regularization," in *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM)*, Hong Kong, China, Feb. 2011, pp. 287–296.
- [17] P. Mittal, C. Papamanthou, and D. Song, "Preserving Link Privacy in Social Network Based Systems," in *Proceedings of the 20th Annual Network and Distributed System Security Symposium (NDSS)*, San Diego, CA, USA, Feb. 2013, pp. 1–16.
- [18] S. J. Murdoch, "Quantifying and Measuring Anonymity," in *Proceedings of the 8th International Workshop on Data Privacy Management and Autonomous Spontaneous Security (DPM)*. Egham, UK: Springer LNCS, volume 8247, Sep. 2013, pp. 3–13.
- [19] A. Narayanan and V. Shmatikov, "Robust De-Anonymization of Large Sparse Datasets," in *IEEE Symposium on Security and Privacy*. Oakland, CA, USA: IEEE, May 2008, pp. 111–125.
- [20] ———, "De-anonymizing Social Networks," in *IEEE Symposium on Security and Privacy*. Oakland, CA, USA: IEEE, May 2009, pp. 173–187.
- [21] P. Pedarsani and M. Grossglauser, "On the Privacy of Anonymized Networks," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '11. San Diego, California, USA: ACM, Aug. 2011, pp. 1235–1243.
- [22] A. Sala, X. Zhao, C. Wilson, H. Zheng, and B. Y. Zhao, "Sharing Graphs Using Differentially Private Graph Models," in *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement conference (IMC)*, Berlin, Germany, Nov. 2011, pp. 81–98.
- [23] K. Sharad and G. Danezis, "An Automated Social Graph De-anonymization Technique," in *Proceedings of the 13th Workshop on Privacy in the Electronic Society (WPES)*. Scottsdale, Arizona, USA: ACM, 2014, pp. 47–58.
- [24] R. Shokri, G. Theodorakopoulos, J. Y. Le Boudec, and J. P. Hubaux, "Quantifying Location Privacy," in *IEEE Symposium on Security and Privacy (S&P)*, Oakland, CA, USA, May 2011, pp. 247–262.
- [25] M. Srivatsa and M. Hicks, "Deanonymizing Mobility Traces: Using Social Network as a Side-Channel," in *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, Raleigh, NC, USA, Oct. 2012, pp. 628–637.
- [26] P. Syverson, "Why I'm Not an Entropist," in *Proceedings of the 17th International Workshop on Security Protocols*. Cambridge, UK: Springer LNCS, volume 7028, Apr. 2009, pp. 213–230.
- [27] B. Thompson and D. Yao, "The Union-Split Algorithm and Cluster-Based Anonymization of Social Networks," in *Proceedings of the 4th International Symposium on Information, Computer, and Communications Security (ASIACCS)*, Sydney, Australia, Mar. 2009, pp. 218–227.
- [28] C. Tofallis, "Add or Multiply? A Tutorial on Ranking and Choosing with Multiple Criteria," *INFORMS Transactions on Education*, vol. 14, no. 3, pp. 109–119, May 2014.
- [29] E. Triantaphyllou, *Multi-Criteria Decision Making Methods: A Comparative Study*. Boston, MA: Springer US, 2000.
- [30] I. Wagner, "Evaluating the Strength of Genomic Privacy Metrics," *ACM Transactions on Privacy and Security (TOPS)*, vol. 20, no. 1, pp. 2:1–2:34, Jan. 2017.
- [31] ———, "Measuring Privacy in Vehicular Networks," in *Proceedings of the 42nd IEEE Conference on Local Computer Networks (LCN)*, Singapore, Oct. 2017, pp. 183–186.
- [32] I. Wagner and D. Eckhoff, "Technical Privacy Metrics: A Systematic Survey," *ACM Computing Surveys*, vol. 51, no. 3, pp. 57:1–57:46, Apr. 2018.

- [33] I. Wagner and Y. Zhao, "Privacy metrics for graph anonymization and de-anonymization," CodeOcean, <http://doi.org/cxv4>, December 2018.
- [34] L. Yartseva and M. Grossglauser, "On the Performance of Percolation Graph Matching," in *Proceedings of the 1st ACM Conference on Online Social Networks*, Boston, MA, USA, Oct. 2013, pp. 119–130.
- [35] X. Ying and X. Wu, "Randomizing Social Networks: a Spectrum Preserving Approach," in *Proceedings of the 2008 SIAM International Conference on Data Mining*, Atlanta, GA, USA, Apr. 2008, pp. 739–750.
- [36] Y. Zhao and I. Wagner, "On the Strength of Privacy Metrics for Vehicular Communication," *IEEE Transactions on Mobile Computing*, 2018.
- [37] ——, "POSTER: Evaluating Privacy Metrics for Graph Anonymization and De-anonymization," in *ASIA CCS '18: 2018 ACM Asia Conference on Computer and Communications Security*. Incheon, Republic of Korea: ACM, June 2018, pp. 817–819.