Influencers and the Giant Component: The Fundamental Hardness in Privacy Protection for Socially Contagious Attributes

Aria Rezaei
arezaei@cs.stonybrook.edu
Stony Brook University

Jie Gao
jg1555@rutgers.edu
Rutgers University

Anand D. Sarwate
anand.sarwate@rutgers.edu
Rutgers University

Abstract

The presence of correlation is known to make privacy protection more difficult. We investigate the privacy of *socially contagious* attributes on a network of individuals, where each individual possessing that attribute may influence a number of others into adopting it. We show that for contagions following the Independent Cascade model there exists a giant connected component of infected nodes, containing a constant fraction of all the nodes who all receive the contagion from the same set of sources. We further show that it is extremely hard to hide the existence of this giant connected component if we want to obtain an estimate of the activated users at an acceptable level. Moreover, an adversary possessing this knowledge can predict the real status ("active" or "inactive") with decent probability for many of the individuals *regardless* of the privacy (perturbation) mechanism used. As a case study, we show that the Wasserstein mechanism, a state-of-the-art privacy mechanism designed specifically for correlated data, introduces a noise with magnitude of order $\Omega(n)$ in the count estimation in our setting. We provide theoretical guarantees for two classes of random networks: Erdős-Rényi graphs and Chung-Lu power-law graphs under the Independent Cascade model. Experiments demonstrate that a giant connected component of infected nodes can indeed appear in real-world networks and a simple inference attack can reveal the status of a good fraction of nodes.

1 Introduction

Protecting the privacy of sensitive user data has always been a major challenge in the big data regime. Without it, users might be reluctant to share honest statistics with service providers and the integrity of collected data can suffer as a consequence. Recent findings have shown that the existence of correlation in data makes privacy protection an even more challenging task. Here, we focus at a special class of correlated data: attributes that spread through a network of individuals by jumping from one person to another. On top of contagious diseases that spread within communities, many behavioral attributes, such as political affiliation, obesity, or smoking habits, are also considered to be *socially contagious* [5]. For an example of the privacy concerns, consider the following scenario: the local Department of Health (DoH) would like to know the fraction of residents who are regular smokers, by conducting a survey from the residents. To protect the privacy of the residents, each resident would report with a probabilistic perturbation (e.g., with a chance of $1/2$ report a random answer). The DoH collects the perturbed reports and calculates the aggregate fraction by removing the introduced bias. If an adversary has the knowledge of the social ties between the participants as well as the social influence model on smoking, how confidently can she guess whether a given individual is a smoker or not? This can be phrased as a statistical inference attack.

It is already recognized that Differential Privacy (DP), arguably the most commonly used privacy framework, along with other conventional privacy frameworks, claim no protection against additional outside information, including knowledge about correlation among data records. In our prior work [19], we have shown using experiments that there is a positive correlation between *out-degree* of a node, which represents the total influence that a node exerts over others, and its vulnerability to an inference attack.

This paper improves upon prior work in two ways. First, we show that obtaining good utility of estimating the fraction of infected users is impossible for a broad class of privacy mechanisms formulated by the success of statistical inference attacks. Secondly, our prior work provided only empirical evidence suggesting a relationship between node's influence and the loss of privacy. In this paper we provide a rigorous explanation of this relationship. To the best of our knowledge, we provide first analysis of the fundamental challenges that exist in data privacy when the attributes are socially contagious and the contagion follows one of the most widely used models, the Independent Cascade model.

Our Contribution: At the core of privacy protection, there is a trade-off between privacy and utility, char-

acterized by the magnitude of noise added to data or query answer. We show that under reasonable assumptions, finding a good trade-off between utility and privacy for such individuals could be extremely difficult. Our contributions are summarized as follows:

- For Erdős-Rényi and Chung-Lu power-law networks [6], we show in Theorems 4.1 and 4.2 that under reasonable assumptions there exists one giant connected component, containing a constant fraction of all nodes, with either all of its nodes possessing the contagious attribute or none. We next show that if an adversary has access to published estimate of the number of individuals with the target attribute, she can learn about the status of this giant component. Using this, she can infer the status of individuals in the network. Those that are most likely to be in this giant component have great influence over the activation of others and are called "influencers". We show that the influencers are most vulnerable to such an inference attack.

- We show in Theorem 5.1 that any privacy-preserving mechanism, in order to increase *uncertainty* about the status of the giant component, needs to introduce an enormous amount of noise (on the order of the size of the network), rendering low data utility.

- We show in Theorem 5.2 that the Wasserstein mechanism [20], a state-of-the-art mechanism that is designed specifically to hide sensitive information in correlated data, will always add noise with a magnitude equal to a constant fraction of the number of all nodes, thus providing weak utility in all cases. This is regardless of the privacy requirements of the application and hence an even stronger result.

- Through experiments on 4 real-world networks, we show that the giant component indeed appears in real networks. Moreover, a sizeable number of nodes are in fact very likely to be a part of it and as a result, an adversary can easily infer their real activation status.

Remark that we go beyond the known results on the difficulties of correlated data privacy under conventional frameworks, such as DP. We make minimal assumptions about the privacy mechanism in question and show that preventing an adversary from inferring the true status of certain nodes is impossible even for very weak utility guarantees. Our results are thus applicable to a wide range of privacy mechanisms, including DP. To see omitted material, check out the full version[1].

---

[1] https://arxiv.org/abs/2012.11877

2 Related Work

Protecting the privacy of correlated data has been a longstanding challenge in the literature. One of the most commonly used privacy frameworks is Differential Privacy (DP). Under DP, entities protect a data record x against an adversary when publishing sensitive data or answering queries, by bounding the probability of the aggregate statistics being the same, in a database with or without x. [9]. Nevertheless, when data is correlated, an adversary could still infer the status of x, by using data records oter than x [15, 12, 25, 14] (notice that this does not violate the DP guarantee) Relaxed definitions of DP have been proposed, such as group differential privacy [8] or specific models for correlated/dependent data [11, 25, 24]. However, the proposed methods do not look at network data or contagion models as we do here, and instead focus on temporal correlation (e.g. location, time series). The most closely related relaxation is the Pufferfish framework [15], where privacy is specifically defined for given pairs of "secrets" which should be *indistinguishable*, and a set of plausible priors which can contain known correlation structures in the data. Specialized Pufferfish mechanisms have been proposed for some problems [15, 23, 12]. For general Pufferfish problems, Song *et al.* [20] proposed the Wasserstein Mechanism which in the end adds noise to a query's answer, with a variance tuned to the assumed prior distribution's structure. We discuss the Wasserstein mechanism in depth in Section 5.

Our work is also different from prior work that consider the network structure to be private and the goal is to report network statistics such as degree distributions [21]. The most relevant to our work in the one by Song *et al.* [20] and Rezaei *et al.* [19]. The former assumes correlated attributes in a social network where nodes of the same community have the same attributes; a rather simplistic model for correlation compared to the widely used contagion models such as Independent Cascade or Linear Threshold. Our work and the one in [19] both consider contagion models, but the results here are more comprehensive, and have stronger theoretical guarantees.

3 Preliminaries

3.1 Data Model Suppose that there are n individuals, each having a binary attribute $x_i \in \{0, 1\}$. Suppose further that the individuals influence one another through a network, $G(V, E)$, where each node $v \in V$ represents an individual and the existence of an edge between two nodes $v, u \in V$ means that the corresponding individuals exert influence on each other. We assume that the attribute $\{x_i\}$ is *contagious*, meaning that starting from a *seed set* of nodes, S ($|S| = s$), the at-

tributes follow a contagion process and spread through the network. The attributes are considered to be sensitive information. In the following, we formally define the contagion process and present the privacy problem statement.

3.2 Contagion Model One of the most commonly used models for contagions is the Independent Cascade (IC) model, used predominantly to model the spread of disease in communities. Let S denote the initial set of nodes that are activated before a contagion starts. Upon activation, each node has exactly one chance to activate each of its neighbors with probability equal to a *transmission rate*, $q \in (0,1]$. The cascade stops when no additional node is activated. Similar to some of the previous works [1], when an edge exists between nodes v and u, we couple the event of v activating u to the event of u activating v.

To analyze cascades over networks, Kempe et al. [13] have introduced the use of *Triggering Set* technique. Suppose that at time 0 we flip a biased coin with probability q for each edge in the graph and choose to "trigger" the edge or not. After that, a node v is activated if and only if there is a path between v and one of the seed nodes via triggered edges. In the case of IC, it suffices for a seed node to appear in the connected component of v for v to become activated.

3.3 Privacy and Utility We consider applications where the number of activated nodes in a population of n individuals, $X_n = \sum x_v$, is of interest, with $\mathcal{X}_n = \{0, 1, \ldots, n\}$ the possible values of X_n. We drop the n when it is clear from context. Since individuals may be reluctant to report their true attributes, an estimate or approximation of X is made using a privacy-preserving mechanism $\mathcal{M}$ which, given the sensitive value X, produces a *perturbed* value $\mathcal{M}(X)$ to report. We measure the utility of $\mathcal{M}$ by the amount of noise it adds to the real value, $|\mathcal{M}(X) - X|$. We associate high utility with relatively low error, and precisely *sublinear asymptotic error*, defined below.

DEFINITION 1. (SUBLINEAR ASYMPTOTIC ERROR)
Consider a sequence of randomized privacy mappings $\mathcal{M}_n \colon \mathcal{X}_n \mapsto \mathcal{X}_n$, where n is the total number of individuals. Then $\mathcal{M}_n$ guarantees sublinear asymptotic *error, if with high probability*[2] *(w.h.p) over $\mathcal{M}$ and the graph generation process we have:*

$$(3.1) \qquad \lim_{n \to \infty} \frac{|\mathcal{M}_n(X_n) - X_n|}{n} \to 0,$$

for all values of $X_n \in \mathcal{X}_n$.

---

[2]E happens with *high* probability if $\lim_{n \to \infty} P(E) \to 1$.

Note that the above indicates that $|\mathcal{M}_n(X_n) - X_n| = o(n)$ for all values of X_n. The association between high utility and sublinear asymptotic error is based on real-world applications and is common in the literature [8]. In many cases, such as when a survey is conducted from a subset of a large population of size n, the margin of error of the resulting estimate from the sample population will be $o(n)$, in other words *negligible* in comparison to the population size, n. Hence, a privacy mechanism that adds noise to X that is, w.h.p, at most $o(n)$ is considered to have high utility since it does not change the value of X in any significant way.

4 Giant Components and the Privacy of Influencers

In this section, we investigate the hardness of protecting the privacy of nodes whose activation causes a large number of other nodes' activation, or *influencers*. Specifically, we will show that over two classes of random networks, if the privacy mechanism guarantees sublinear asymptotic error (from Definition 1), one can infer the status of the influencers with high confidence regardless of the details of the privacy mechanism and the amount of noise added. The specific definition of influencers will depends on the characteristics of the graph. For the two random networks that we study, influencers are simply those with highest degrees. In both networks, and in real graphs, under the right circumstances a *giant component* consisting of a constant fraction of all nodes appears in the triggering set of the network. If any of the seed nodes end up in this giant component, all nodes inside of it will be activated. We show that the knowledge of the activation of such a cluster of nodes allows an adversary mount a successful inference attack.

4.1 Erdős-Rényi Graphs An *Erdős-Rényi* graph is an undirected and simple graph $G(V, E)$ with n nodes ($|V| = n$) where each possible edge (v, u) exists with probability $p \in (0, 1)$.

In the case where $np > 1$, the size of the largest connected component, known as the *giant* component, is $\Theta(n)$ w.h.p. More specifically, suppose that C_i is the ith largest connected component and $|C_i|$ denotes its size. Then, if $np > 1$, $|C_1| \sim yn$ where y is the solution to the following [2]:

$$(4.2) \qquad e^{-npy} = 1 - y.$$

Furthermore, all other components are *tiny*; for all $i \geq 2$, we have $|C_i| = O(\log n)$. Finally, in case where np is a constant above 1, w.h.p the maximum degree is [10]:

$$(4.3) \qquad \Delta \approx \frac{\log n}{\log \log n},$$

Note that the triggering set of an Independent Cascade with transmission rate q over $G(n,p)$ is itself an Erdős-Rényi graph $G(n,pq)$. This means that if $npq > 1$, a connected component with a fraction of all nodes exists; either nodes inside the cluster are all activated, or none of them are.

We now present our main theorem on the hardness of protecting privacy for influencers. We denote the triggering set created by the process outlined in Section 3.2 by H and its ith largest connected component by C_i^H.

THEOREM 4.1. *Suppose that the nodes in $G(n,p)$ are numbered in the decreasing order of their degree; with node i having the i-th highest degree. Suppose further that the number of seeds $s = o(n/\log^2 n)$, the privacy mechanism $\mathcal{M}$ with maximum error $e_{\mathcal{M}}$ guarantees sublinear asymptotic error, and $npq > 1$. Then, for any $\epsilon \in [0,1)$, there exists a $k \geq 1$ such that:*

$$\Pr[\mathrm{x}_i = 1 \mid \mathcal{M}(X) \geq |C_1^H| - e_{\mathcal{M}}] \geq 1 - \epsilon$$
$$(4.4) \quad \Pr[\mathrm{x}_i = 0 \mid \mathcal{M}(X) \leq s|C_2^H| + e_{\mathcal{M}}] \geq 1 - \epsilon,$$

for all $i \leq k$ and large enough n.

Here, we have assumed that the number of initial seeds is not an unrealistically high number, as is usually the case in real-world scenarios. For instance, organic cascades, such as a contagious disease, originate from a handful of initial infected individuals. We have also assumed that each node, upon activation, activates *on average* at least one other node ($npq > 1$).

Notice that the two conditions in (4.4) correspond to when C_1^H is activated and $\mathcal{M}(X) \geq |C_1^H| - e_{\mathcal{M}}$, or when C_1^H is inactive and $\mathcal{M}(X) \leq e_{\mathcal{M}} + s|C_2^H|$. Since $npq > 1$, w.h.p $C_1^H \sim yn$ where y is the solution to (4.2) when p is substituted with pq. This means that the two conditions in (4.4) are also exhaustive since the difference between the two values on the right side of the inequalities is:

$$|C_1^H| - s|C_2^H| - 2e_{\mathcal{M}} = \Theta(n).$$

This means that one could infer the activation of C_1^H from the value of $\mathcal{M}(X)$ with with high confidence. We make this claim more rigorous in Section 5.

As an example, consider the case when $n = 10,000$, $s = \log n \leq 10$, $|\mathcal{M}(X) - X| \leq \sqrt{n}$, and $y = 0.4$ ($pq \approx 1.28$). In the event that C_1^H is not activated we will have $\mathcal{M}(X) \leq \log^2(10,000) + 100 \leq 200$, while in the event of C_1^H's activation we have $\mathcal{M}(X) \geq 3900$. Since the value of $\mathcal{M}(X)$ reveals the activation status of C_1^H, we can use this information to bound $\Pr[\mathrm{x}_i = \cdot]$ using the two lemmas below the proof of which you can find in Section B (see full version).

LEMMA 4.1. *With a given $\mathcal{M}(X)$, for node i we have:*

(1) $\Pr[\mathrm{x}_i = 1 \mid \mathcal{M}(X) \geq |C_1^H| - e_{\mathcal{M}}] \geq \Pr[i \in C_1^H]$

(2) $\Pr[\mathrm{x}_i = 0 \mid \mathcal{M}(X) \leq s|C_2^H| + e_{\mathcal{M}}] \geq \Pr[i \in C_1^H]$.

LEMMA 4.2. *When n is large enough, the probability that a node i with a fixed degree k in H is in C_1^H is approximated by:*

$$(4.5) \quad \Pr[i \notin C_1^H] \sim -\exp(-ky).$$

The degree of any node i in H, denoted by d_i^H, follows a binomial distribution characterized by d, its degree in G, and q, the transmission rate. Applying Chernoff bound for lower tail, we have:

$$(4.6) \quad \Pr\left[d_i^H \leq \frac{dq}{2}\right] \leq -\exp\left(-\frac{dq}{8}\right).$$

Since the function $\exp(-ky)$ is decreasing in k, we have:

$$(4.7) \quad \Pr[i \notin C_1^H] \leq \exp(0)\Pr\left[d_i^H \leq \frac{dq}{2}\right] + \exp\left(-\frac{dqy}{2}\right)\Pr\left[d_i^H > \frac{dq}{2}\right].$$

Combining this with (4.6), we have:

$$(4.8) \quad \Pr[i \notin C_1^H] \leq \exp\left(-\frac{dq}{8}\right) + \exp\left(-\frac{dqy}{2}\right).$$

Plugging in the above node 1 with maximum degree, $\Delta \approx \log n/\log\log n$ from (4.3), we can write for any $\epsilon > 0$ and an n large enough:

$$(4.9) \quad \Pr[1 \in C_1^H] \geq 1 - \epsilon,$$

This satisfies the conditions in (4.4) according to Lemma 4.1. Depending on the value of ϵ, there will be other high-degree nodes for which the bound in (4.8) goes below ϵ since its value decreases exponentially in d. This shows that the k in Theorem 4.1 is always ≥ 1 which concludes the proof of Theorem 4.1.

4.2 Chung-Lu Power-Law Networks Despite being extensively studied, Erdős-Rényi graphs do not represent many phenomena observed in reality; in particular, the *power-law* degree distribution that is common in many real-world networks [3]. To study such networks, we turn to *Chung-Lu* networks [6] with a degree distribution that follows a power-law distribution. Let $\langle w_1, \ldots, w_n \rangle$ ($w_i \in \mathbb{R}_+$) be weights assigned to n individuals and $\ell_n = \sum w_i$. A *Chung-Lu* network is comprised of n nodes, where an edge between two nodes i and j appears with probability:

$$(4.10) \qquad P_{ij} = \min\left(1, \frac{w_i w_j}{\ell_n}\right).$$

To generate a network that resembles a power-law network, we assign the weights using the following function:

$$(4.11) \qquad w_i = [1 - F]^{-1}(i/n), \quad F(x) = 1 - (d/x)^b,$$

where b determines the *scale* of the power-law distribution and d is the minimum expected degree in the graph. You can see that the i-th ranked node has a weight corresponding to the i/n percentile of the distribution. Finding the solution to (4.11) yields:

$$(4.12) \qquad w_i = d\left(\frac{n}{i}\right)^\beta, \quad \beta = \frac{1}{b}.$$

Note that the triggering set of a Chung-Lu network with parameters d and b for an IC with parameter q is itself a Chung-Lu network with parameters dq and b. It has been shown [1, 22] that if (1) $b \in (0,2]$ or (2) $b > 2$ and $d > (b-1)(b-2)$, the generated graph will have a giant component with a constant fraction of nodes, αn ($0 < \alpha < 1$), and many small components each being at most $O(\log n)$ in size.

As before, we denote the main graph by G, its triggering set by H, and the i-th biggest connected component of H by C_i^H.

THEOREM 4.2. *Suppose that $s = o(n/\log^2 n)$, the privacy mechanism $\mathcal{M}$ with maximum error $e_{\mathcal{M}}$ guarantees sublinear asymptotic error, and either (1) $b \in (0,2]$ or (2) $b > 2$ and $dq > (b-1)(b-2)$. Then, for any $\epsilon > 0$, there is a subset of nodes nodes $V' \subseteq V$ that grows arbitrarily large in size ($|V'| = \omega(1)$) and for any node $i \in V'$, we have:*

$$\Pr[\mathrm{x}_i = 1 \mid \mathcal{M}(X) \geq |C_1^H| - e_{\mathcal{M}}] \geq 1 - \epsilon$$
$$(4.13) \qquad \Pr[\mathrm{x}_i = 0 \mid \mathcal{M}(X) \leq s|C_2^H| + e_{\mathcal{M}}] \geq 1 - \epsilon,$$

The proof of this theorem follows arguments similar to Theorem 4.1 and is available in full in Section C (see full version). Depending on the value of b, there could be 3 upper-bounds for the value of i, all 3 of which can grow arbitrarily in size:

$$(4.14) \qquad 0 < b < 1: \; i = o\left(n^b\right),$$

$$(4.15) \qquad b = 1: \; i = o\left(\frac{n}{\log n}\right)$$

$$(4.16) \qquad b > 1: \; i = o(n).$$

The 3 equations above show that for any ϵ in Theorem 4.2, the subset $V' \subseteq V$ includes all nodes ranked
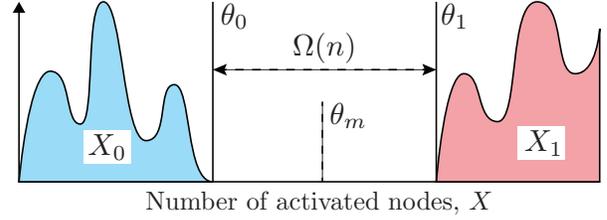

Figure 1: Probability distribution of X, when C_1^H is *not* activated, X_0 (left, blue), with maximum value θ_0, and when C_1^H *is* activated, X_1 (right, red), with minimum value θ_1. Note that $\theta_1 - \theta_0 = \Omega(n)$ in our setting.

$i \leq f(n)$ where f can be any increasing function of n that meets the corresponding asymptotic condition depending on b, for instance $f(n) = \log^2 n$ in the case of $b = 1$.

4.3 Notes on General Graphs In general graphs too, if one retains each edge of a connected graph with probability p, for values of p above a certain threshold[3] a *giant component* appears with a constant fraction of all nodes and all the other components are of size $o(n)$ [7]. This means that our findings are applicable to general graphs as well, however, the vulnerability of each node depends on its probability of appearance in that giant component, which depends on the structure of the graph. We show through experiments in Section 6 that for reasonable values of the transmission rate, q, a giant component appears in the triggering set and a sizeable number of all the nodes frequently appear in that component and are hence vulnerable to the inference attack introduced in this section.

5 Hiding the Status of the Giant Connected Component

As we saw in Section 4, when a privacy mechanism fails to hide the status of the giant component, C_1^H, an adversary can infer the real status of individual nodes by estimating the probability of their presence in C_1^H. Here, we show that if there is to be *any uncertainty* about the activation status of C_1^H, then the privacy mechanism *cannot* provide sublinear asymptotic error. We denote by X_1 and X_0 the distribution of X when C_1^H is active and inactive, respectively. Let $\mathcal{M}: [n] \mapsto [n]$ be a randomized privacy mechanism. Then, we denote the distribution of *perturbed* values for X_0 and X_1 by $Z_0 = \mathcal{M}(X_0)$ and $Z_1 = \mathcal{M}(X_1)$. For brevity, we denote the probability density functions of these distributions

---

[3]Define the second-order average degree $\tilde{d} = \sum_v d_v^2/(\sum_v d_v)$ where d_v is the degree of v. If $p > (1 + \varepsilon)/\tilde{d}$, then almost surely, the percolated subgraph has a giant component. Otherwise, if $p < (1 - \varepsilon)/\tilde{d}$, almost surely, there would be no giant component.

by the same symbol, i.e. $\Pr[\mathcal{M}(x \sim X_0) = a] = Z_0(a)$. We frame the problem as a hypothesis testing scenario. Suppose that the reported value z is observed. Then, the two hypotheses are: (1) H_0, z comes from Z_0 and (2) H_1, z comes from Z_1.

THEOREM 5.1. *If the hypothesis testing error is at least a constant $0 < \epsilon < 1$, then $\mathcal{M}$ cannot provide sublinear asymptotic error.*

Proof. Here, we make use of the relation between hypothesis testing error and the total variation distance (TVD), calculated between two distributions μ and ν over a range $\mathcal{X}$ as below:

$$(5.17) \qquad \text{TVD}(\mu, \nu) = \frac{1}{2} \sum_{x \in \mathcal{A}} |\mu(x) - \nu(x)|.$$

Since the hypothesis testing error is $1 - \text{TVD}(Z_0, Z_1)$ and given that it is at least ϵ, for the total variation distance between Z_0 and Z_1 we have:

$$(5.18) \quad \text{TVD}(Z_0, Z_1) = \frac{1}{2} \sum_a |Z_1(a) - Z_0(a)| \le 1 - \epsilon.$$

Suppose that:

$$\theta_0 = \max X \sim X_0, \ \theta_1 = \min X \sim X_1,$$

and $\theta_m = (\theta_0 + \theta_1)/2$. Figure 1 depicts the relation between the two distributions X_0 and X_1, and the values θ_0, θ_1 and θ_m. We can rewrite (5.18) as:

$$\sum_{a \le \theta_m} Z_1(a) - Z_0(a) + \sum_{a > \theta_m} Z_0(a) - Z_1(a) \le 1 - \epsilon$$

$$(5.19) \qquad \sum_{a \le \theta_m} Z_1(a) + \sum_{a > \theta_m} Z_0(a) \ge \epsilon.$$

Thus, at least one of the terms on the left side of (5.19) is greater than or equal to $\epsilon/2$. Without loss of generality, we assume that it's $\sum Z_1(a)$. We then have:

$$(5.20) \quad \sum_{a \le \theta_m} Z_1(a) = \sum_b X_1(b) \Pr[\mathcal{M}(X) \le \theta_m] \ge \frac{\epsilon}{2}.$$

Assuming that $X^* = \arg \max_{X \sim X_1} \Pr[\mathcal{M}(X) \le \theta_m]$, since $\sum_b X_1(b) = 1$, we have:

$$(5.21) \qquad \Pr[\mathcal{M}(X^*) \le \theta_m] \ge \frac{\epsilon}{2}.$$

Since $\theta_1 - \theta_m = \Omega(n)$, we have at least one $X \sim X_1$ (X^*) for which *with constant probability* the perturbed value, $\mathcal{M}(X^*)$, is $\Omega(n)$ distance away. This means that $\mathcal{M}$ cannot provide Sublinear Asymptotic Error. $\square$

5.1 Wasserstein Mechanism's Utility As a case study on the hardness of providing privacy in our setting, we turn to the Wasserstein mechanism [20]; a state-of-the-art privacy mechanism designed specifically for correlated data privacy. This mechanism is based on the Pufferfish framework [15]. Similar to DP, Pufferfish tries to make pairs of *secrets* indistinguishable up to a factor e^ε. For instance, here we need to make it hard for an adversary to distinguish between pairs of secrets: "$x_i = 1$" and "$x_i = 0$" where $i \in V' \subseteq V$. A key difference between Pufferfish and DP is that in Pufferfish, these pairs of secrets can be *non-exhaustive*, meaning that one can decide to protect only a subset of all individuals. A more detailed definition of the Pufferfish framework and its application in our setting is available in Section A (see full version).

Here, we show that in a setting similar to ours, for *any* subset of individuals that the Wasserstein mechanism intends to protect, with reasonable utility requirements, the magnitude of the added noise added will be $\Omega(n)$. Note that this result is stronger than our previous findings for more general privacy mechanisms in Section 4, as there is no inference attack involved and the mechanism adds $\Omega(n)$ noise to protect nodes with even little chance of being in the giant component.

The Wasserstein mechanism is based on the ∞-Wasserstein distance between two distributions. Suppose that μ and ν are two probability distributions on $\mathbb{R}$ and denote the set of all joint distributions with marginals μ and ν by $\Gamma(\mu, \nu)$. Then, the ∞-Wasserstein distance, or in short W_∞, is calculated as below:

$$(5.22) \qquad W_\infty(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu, \nu)} \max_{x,y, \gamma(x,y) > 0} |x - y|.$$

If you consider the probability mass of μ to be piles of dust along $\mathbb{R}$ and ν as the distribution of tiny holes with different depths, you can think of any $\gamma \in \Gamma(\mu, \nu)$ as a plan to move the piles of dust into the holes and W_∞ as the minimum possible "maximum distance any speck of dust travels from a pile in μ into a hole in ν".

Given $\mathbb{Q}$, the set of all pairs of secrets and a sensitive query F, the Wasserstein Mechanism adds noise via $\text{Lap}(W/\varepsilon)$ to achieve ε-Pufferfish privacy where W is:

$$(5.23) \qquad W = \sup_{(s_i, s_j) \in \mathbb{Q}} W_\infty(\mu_i, \mu_j),$$

where $\mu_i = \Pr[F(X) = \cdot \mid s_i]$. Song et al. have proven that the Wasserstein Mechanism will always add less noise than *group* differential privacy [20], which makes our findings applicable to group differential privacy as well.

THEOREM 5.2. *Suppose that C_i^H is the i-th biggest connected component of the triggering set, H, of a*

connected graph G. Suppose further that X is the total number of activated nodes and $\mathcal{W}(W, \varepsilon)$ is a Wasserstein Mechanism that adds noise to X following $Lap(W/\varepsilon)$ to provide ε-Pufferfish for a subset $V' \subseteq V$ of nodes. If we have: (1) $|C_1^H| = \Theta(n)$ and (2) $s|C_2^H| = o(n)$, then $W = \Omega(n)$ for any $V' \subseteq V$.

Proof. For a node v, suppose that $\mu_i^v = P(X = \cdot \mid x_v = i)$ for $i \in \{0, 1\}$. Similar to Theorem 5.1, we denote by θ_0 the *largest* cascade when C_1^H is *not* activated, and by θ_1 the *smallest* cascade when C_1^H *is* activated. Note that $\theta_1 - \theta_0 = \Omega(n)$ according to conditions (1) and (2). Then, the probability of $X \leq \theta_0$ when v is activated is:

$$(5.24) \quad \mu_1(X \leq \theta_0) = \Pr[v \notin C_1^H]\Pr[S \cap C_1^H = \emptyset].$$

C_1^H should not receive any seed nodes, and v, which is activated should not be in C_1^H. For the probability of $X \leq \theta_0$ when v is *not* activated we have:

$$(5.25) \quad \mu_0(X \leq \theta_0) = \Pr[v \in C_1^H] + \\ \Pr[v \notin C_1^H]\Pr[S \cap C_1^H = \emptyset].$$

Let $\mathcal{A} = [0, \theta_0] \times [\theta_1, n]$. For any joint distribution γ with marginals μ_0 and μ_1, we have:

$$(5.26)$$
$$\int_{(x,y)\in\mathcal{A}} \gamma(x,y)d(x,y) \geq |\mu_0(X \leq \theta_0) - \mu_1(X \leq \theta_0)|$$
$$(5.27) \qquad\qquad \geq \Pr[v \in C_1^H].$$

(5.26) is due to the fact that any difference in the total mass of probability in the region $X \leq \theta_0$ between μ_0 and μ_1 has to be transferred into the region $X \geq \theta_1$. Due to (5.24) and (5.25), this difference is equal to $\Pr[v \in C_1^H]$. Since G is connected, for any v: $\Pr[v \in C_1^H] > 0$, which yields:

$$(5.28) \qquad \exists\, x \in [0, \theta_0], y \in [\theta_1, n]: \gamma^*(x, y) > 0,$$

resulting into:

$$W_\infty(\mu_0, \mu_1) \geq \theta_1 - \theta_0 = \Omega(n).$$

Since W is selected by taking the maximum W_∞ across all $v \in V'$, we have $W = \Omega(n)$. $\square$

Note that conditions (1) and (2), as we discussed in Section 4, can be satisfied both in the random networks we studied and in real-world networks. With $W = \Omega(n)$, for reasonable values of ε, such as a constant, the added noise to protect *any* subset of nodes follows $Lap(\Omega(n))$ distribution, which destroys data utility.

It is worth noting that the underlying reason for the massive difference between μ_1^v and μ_0^v for any node v is indeed the disparity between the values of X when C_1^H is activated and when it is not.
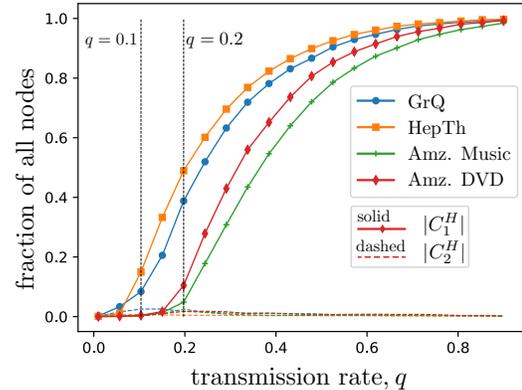

Figure 2: The average size of the largest, C_1^H in solid marked lines, and the second largest connected components, C_2^H in dotted lines, of the triggering set as fraction of all nodes, with with different values of q.

6 Experiments

Our experiments on real-world data sets show that giant connected components appear in cascades over real networks under reasonable values of the transmission rate q. Moreover, a sizeable number of nodes in these networks appear frequently in the giant connected component and are highly vulnerable to inference attacks since hiding the giant component is difficult.

We conduct our experiments on 4 real-world networks (Table 1): two co-authorship networks, GrQc and HepTh [17] and two co-purchase networks, Amazon DVDs and Amazon Musics [16]. The last two columns of this table includes the average size of the giant component ($|\overline{C_1^H}|$) and the second largest component ($|\overline{C_2^H}|$) of the triggering set, H, when q is 0.3. For each net-

Table 1: Real-world networks.

Network	#Nodes	#Edges	$|\overline{C_1^H}|$	$|\overline{C_2^H}|$
GrQc	2.4k	10.9k	1,577.3	28.99
HepTh	4.9k	19.4k	3,494.7	19.73
Amz. Music	29.4k	84.3k	9,804.0	197.23
Amz. DVDs	9.5k	27.6k	4,341.5	139.39

work, 1000 triggering sets are generated at random. As a result of the massive difference seen between $|C_1^H|$ and $|C_2^H|$, it will be extremely hard to hide the activation status of C_1^H. To find out for what values of q a giant connected component appears, we have tested 20 values of q between 0.05 and 0.9 and reported $\overline{C_1^H}$ and $\overline{|C_2^H|}$ over 50 realizations. The results are depicted in Figure 2. Initially, for extremely low values of q, the two are almost indistinguishable in size, but as q grows, C_1^H quickly transitions into the *giant component*, contain-

Table 2: Number and percentage of nodes with probability of being in C_1^H higher than different thresholds.

| Network | Threshold for $\Pr[v \in C_1^H]$ | | | | |
	$\geq 99\%$	$\geq 95\%$	$\geq 90\%$	$\geq 75\%$	$\geq 50\%$
GrQc	211 (8.71%)	383 (15.81%)	532 (21.97%)	906 (37.41%)	1,765 (72.87%)
HepTh	462 (9.41%)	966 (19.68%)	1371 (27.93%)	2414 (49.17%)	4054 (82.58%)
Amz. Music	153 (0.52%)	408 (1.39%)	751 (2.56%)	2865 (9.76%)	8506 (28.98%)
Amz. DVDs	58 (0.61%)	203 (2.14%)	414 (4.36%)	1647 (17.36%)	4529 (47.73%)
ER(2500,5)	1 (0.0%)	3 (0.1%)	10 (0.4%)	433 (17.3%)	1,737 (69.5%)
CL(2500,5,1.1)	668 (26.7%)	994 (39.8%)	1,269 (50.8%)	2,043 (81.7%)	2,459 (98.4%)
CL(2500,5,1.5)	952 (38.1%)	1,534 (61.4%)	1,902 (76.1%)	2,354 (94.2%)	2,481 (99.2%)
CL(2500,5,2)	622 (24.9%)	1,171 (46.8%)	1,599 (64.0%)	2,234 (89.4%)	2,446 (97.8%)

ing a sizeable portion of all nodes while $|C_2^H|$ remains negligible in comparison to n. For values of q as low as 0.1 in the case of the two co-authorship networks, we already see a significant difference between the two components. This sharp increase happens in the co-purchase networks, Amazon DVD and Music, at around $q = 0.2$. In comparison, the transmission probability of high-risk contacts in COVID-19 is estimated at about 15.6% [18]. This shows that in real-world contagions there could be a giant connected component that endangers the privacy of those individuals in it. Next, we estimate the probability of each node's presence in C_1^H to measure their vulnerability against inference attacks. The results for the 4 real-world networks, along with 4 random synthetic networks for comparison, are available in Table 2. In each column, we report the number and percentage of nodes whose probability of appearing in C_1^H is above a threshold mentioned at the top. Results are aggregated over 1000 realizations of H for each network, using $q = 0.3$, and the thresholds are 99%, 95%, 90%, 75% and 50% respectively. First, notice that there are hundreds of nodes across all 4 networks with values above even very high thresholds, such as 99% or 95%. This means that if the privacy mechanism has Sublinear Asymptotic Error, an adversary can infer the real status of these nodes with a confidence above 95% or 99%. In addition, the numbers vary between the co-authorship networks, GrQ and HepTh, and the Amazon DVD and Music networks. As discussed earlier, the structure of each network plays a key role in these probability values. For comparison, we repeat this experiment on 4 synthetic randomized networks with 2500 nodes as well: an Erdős-Rényi graph with $n = 2500$ and $np = 5$, and 3 Chung-Lu power-law networks with minimum expected degree, $d = 5$ and the scale parameter, b equal to 1.1, 1.5, and 2 (real-world values lie in $(1, 2]$ [3]). With the selected setting, all 4 of these networks will have a giant component in their corresponding triggering sets w.h.p. As expected, we observe that the real-world networks

are much more resembled by the Chung-Lu power-law networks, rather than the Erdős-Rényi graphs. As we discussed in Section 4.2, Erdős-Rényi graphs do not model the uneven degree distribution of real-world networks, which is partially responsible for the unbalanced distribution of $\Pr[v \in C_1^H]$ among nodes.

7 Conclusion and Future Work

It is known that if a contagion follows the Independent Cascade model, a giant connected component with a constant fraction of all individuals likely appears. We established that if a privacy mechanism provides sublinear asymptotic error, hiding the status of this component is impossible. Because of this, many individuals are vulnerable to inference attacks, with those that greatly influence others, a.k.a influencers, in the most danger. Finally, we have shown that the current state-of-the-art mechanism is incapable of providing sublinear ($o(n)$) privacy for *any* node in the two classes of random networks we studied and in general graphs. Our findings put new emphasis on the importance of protecting the topology of social networks. It is interesting to study other contagion models, such as the Linear Threshold model, seeking for a similar impossibility result. The final question is, could there be adequate privacy for individuals with new privacy frameworks or minimal changes in our assumptions?

References

[1] Mohammad Akbarpour, Suraj Malladi, and Amin Saberi. Just a few seeds more: value of network information for diffusion. *Available at SSRN 3062830*, 2018.

[2] Noga Alon and Joel H Spencer. *The probabilistic*

method. John Wiley & Sons, 2004.

[3] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286 (5439):509–512, 1999.

[4] Edward Chlebus. An approximate formula for a partial sum of the divergent p-series. *Applied Mathematics Letters*, 22(5):732–737, 2009.

[5] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4): 370–379, 2007.

[6] Fan Chung and Linyuan Lu. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6(2):125–145, 2002.

[7] Fan Chung, Paul Horn, and Linyuan Lu. Percolation in general graphs. *Internet Math.*, 6(3):331–347, January 2009.

[8] C. Dwork and A. Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4): 211–407, 2013. doi: 10.1561/0400000042.

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, pages 265–284, 2006.

[10] Alan Frieze and Michał Karoński. *Introduction to random graphs*. Cambridge University Press, 2016.

[11] A. Ghosh and R. D. Kleinberg. Inferential privacy guarantees for differentially private mechanisms. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, 2017.

[12] Xi He, Ashwin Machanavajjhala, and Bolin Ding. Blowfish privacy: Tuning privacy-utility trade-offs using policies. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pages 1447–1458, 2014.

[13] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, 2003.

[14] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, pages 193–204, 2011.

[15] Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)*, 39(1):3, 2014.

[16] Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5–es, 2007.

[17] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2–es, 2007.

[18] Chayanon Phucharoen, Nichapat Sangkaew, and Kristina Stosic. The characteristics of covid-19 transmission from case to high-risk contact, a statistical analysis from contact tracing data. *EClinicalMedicine*, 2020.

[19] Aria Rezaei and Jie Gao. On privacy of socially contagious attributes. In *Proceedings of the 19th IEEE International Conference on Data Mining (ICDM'19)*, pages 1294–1299, November 2019.

[20] Shuang Song, Yizhen Wang, and Kamalika Chaudhuri. In *Proceedings of the 2017 ACM International Conference on Management of Data*, New York, NY, USA.

[21] C Task and C Clifton. A guide to differential privacy theory in social network analysis. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 411–417, August 2012.

[22] Remco Van Der Hofstad. *Random graphs and complex networks*, volume 1. Cambridge university press, 2016.

[23] Bin Yang, Issei Sato, and Hiroshi Nakagawa. Bayesian differential privacy on correlated data. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, SIGMOD '15, pages 747–762, New York, NY, USA, 2015. ISBN 9781450327589. doi: 10.1145/2723372.2747643.

[24] J. Zhao, J. Zhang, and H. V. Poor. Dependent differential privacy for correlated data. In *2017 IEEE Globecom Workshops (GC Wkshps)*, pages 1–7, 2017.

[25] T. Zhu, P. Xiong, G. Li, and W. Zhou. Correlated differential privacy: Hiding information in non-iid data set. *IEEE Transactions on Information Forensics and Security*, 10(2):229–242, 2015.

A Pufferfish Privacy

In Pufferfish privacy, the model seeks to protect pairs of "secrets" by making the two secrets *indistinguishable* from each other by an adversary given the output of the privacy mechanism, $\mathcal{M}(X)$. In our case, the two secrets are whether $x_v = 0$ or $x_v = 1$ for each v. Pufferfish privacy also provide guarantees with respect to a set of *data-generating* distributions governing how the dataset is created. The meaning of the privacy guarantee is then conditional: if the data was generated according to the assumed family of data-generating distributions, then a privacy mechanism will make it hard for an adversary to distinguish whether secret i, e.g. $x_v = 0$, or secret j, e.g. $x_v = 1$, is true given that $\mathcal{M}(X)$ is known by the public. Mechanisms that guarantee Pufferfish privacy introduce noise into the computation to provide this guarantee.

More formally, let $\mathbb{S}$ be a set of secrets, $\mathbb{Q} \subseteq \mathbb{S} \times \mathbb{S}$ a set of pairs of secrets we wish an adversary cannot distinguish between, and Π a set of *data-generating* distributions, $\pi \in \Pi$, governing how the dataset is created. The formal definition of Pufferfish privacy is given below.

DEFINITION 2. (PUFFERFISH PRIVACY [15]) *A randomized mechanism $\mathcal{M}$ in framework ($\mathbb{S}, \mathbb{Q}, \Pi$) provides ε-Pufferfish privacy if for any data-generating distribution $\pi \in \Pi$, any pairs of secrets $s_i, s_j \in \mathbb{Q}$ such that $P(s_i \mid \pi), P(s_j \mid \pi) > 0$, and for all possible outputs $w \in Range(\mathcal{M})$ we have:*

$$\text{(A.1)} \qquad \frac{\Pr[\mathcal{M}(X) = w \mid s_i, \pi]}{\Pr[\mathcal{M}(X) = w \mid s_j, \pi]} \leq e^{\varepsilon}.$$

Note that the above is almost exactly the guarantee seen in Differential Privacy. The key difference is that the randomness of X, the dataset, is taken into consideration when the probability ratio in (A.1) is calculated. The goal of Pufferfish privacy is to incorporate correlations between data points, via the assumptions on Π, to enable us to measure the privacy guarantees for correlated data; something that DP is known to lack. For instance, in our case the set of data-generating distributions could be *all* possible graphs generated by a random network given fixed parameters. Alternatively, with a fixed graph, the set Π could include all possible triggering sets. The flexibility of $\mathbb{S}$ and $\mathbb{Q}$ allows the mechanism designer to only aim for the protection of the privacy of a subset of nodes, not all of them.

B Proofs of Lemmas 4.1 and 4.2

B.1 Lemma 4.1

Proof. As we discussed in Section 4.1, the condition in the first equation means that C_1^H is activated. Once C_1^H is active, one way for node i to become activated is to be a part of C_1^H. Obviously, there are other ways for that to happen, hence the lower bound. An exactly similar argument goes for the second equation as well. □

B.2 Lemma 4.2

Proof. Since we are fixing the degree to be k, and each node has an equal probability of being one of the k neighbors of i, the probability that i chooses *none* of the nodes in C_1^H as a neighbor is equal to:

$$\text{(B.2)} \qquad \frac{\binom{n-m}{k}}{\binom{n}{k}} \sim -\exp\left(\frac{km}{n}\right).$$

In the above, we have used Stirling's approximation of $n! \sim \sqrt{2\pi}(n/e)^n$. Substituting m in (B.2) with $yn \pm o(n)$ as the number of nodes in C_1^H yields (4.5). □

C Proof of Theorem 4.2

Proof. First notice that unlike Theorem 4.1, here we also prove that the subset of nodes that will not have adequate protection by $\mathcal{M}$ can in fact grow arbitrarily in size. In other words:

$$\lim_{n \to \infty} |V'| \to \infty.$$

Similar to the arguments in Theorem 4.1, we can observe that either C_1^H is active and $\mathcal{M}(X) \geq |C_1^H| - e_\mathcal{M} = \Omega(n)$, or it is inactive and $\mathcal{M}(X) \leq s|C_2^H| + e_\mathcal{M} = o(n)$. Because of this, one can infer the status of C_1^H given the value of $\mathcal{M}(X)$. We can now use Lemma 4.1 here to bound the two probabilities in (4.13) by $\Pr[i \in C_1^H]$. For its complement, $\Pr[i \notin C_1^H]$ we have:

$$\Pr[i \notin C_1^H] \leq \prod_{j \in C_1^H} 1 - \min\left(1, \frac{w_i w_j}{\ell_n}\right)$$
$$\text{(C.3)} \qquad \leq \left(1 - \min\left(1, \frac{w_i w_n}{\ell_n}\right)\right)^{\alpha n},$$

since $w_n = dq < w_j$ $(n > j)$ due to (4.12). In extreme cases when b is too small, we might have $w_1 w_n / \ell_n \geq 1$, making the node with the highest degree $(i = 1)$ connected to every single node with probability 1 and the whole graph becomes connected. This means that with any number of seed nodes $s > 0$, the *whole* network is activated and there could be no privacy for any node. Apart from such cases, which are rare in reality as b usually lies in $(1, 2]$ in real-world networks [3], we can

assume that $w_i w_n \leq \ell_n$ and rewrite (C.3) as:

$$\Pr[i \notin C_1^H] \leq \exp\left(-\frac{w_i w_n \alpha n}{\ell_n}\right)$$

$$\leq \exp\left(-\frac{d^2 q^2 \alpha n^{\beta+1}}{i^\beta dq n^\beta \sum_j j^{-\beta}}\right)$$

(C.4) $$\leq \exp\left(-\frac{dq\alpha n}{i^\beta \sum_j j^{-\beta}}\right).$$

The above approaches 0 as n grows large enough when:

(C.5) $$i = o\left(\left[\frac{n}{\sum_j j^{-\beta}}\right]^{1/\beta}\right)$$

Depending on the value of β $(1/b)$, there could be 3 cases for (C.5):

1. In case $\beta > 1$ $(0 < b < 1)$, the summation $\sum_j^\infty j^{-\beta}$ is a constant number. Then, for i we can write:

   (C.6) $$i = o\left(n^{1/\beta}\right) = o\left(n^b\right).$$

2. In case $\beta = 1$, the summation $\sum_{j=0}^n j^{-1}$ tends to $\Theta(\log n)$. Using this, we can write:

   (C.7) $$i = o\left(\frac{n}{\log n}\right)$$

3. In case $0 < \beta < 1$, the partial summation $\sum_{j=0}^n j^{-\beta}$ is $\Theta(n^{1-\beta})$ [4]. Using this, we can write:

   (C.8) $$i = o\left(\left[\frac{n}{n^{1-\beta}}\right]^{1/\beta}\right) = o(n)$$

The 3 equations above show that for any ϵ in Theorem 4.2, the subset $V' \subseteq V$ includes all nodes ranked i where $i \leq f(n)$ and f meets the corresponding asymptotic condition depending on b in the equations above. Note that $f(n)$, and subsequently $|V'|$, can grow arbitrarily large. For instance in the case of $b = 1$, to meet the condition in (C.7), we can set $f(n) = n/\log^2 n$. This concludes the proof. $\qquad \square$