CAMBRIDGE
UNIVERSITY PRESS

RESEARCH ARTICLE

Large very dense subgraphs in a stream of edges*

Claire Mathieu[1] and Michel de Rougemont[2,*] (ID)

[1]CNRS and IRIF, Paris, France and [2]University Paris II and IRIF, Paris, France
*Corresponding author. Email: mdr@irif.fr

Action Editor: Ulrik Brandes

Abstract

We study the detection and the reconstruction of a large very dense subgraph in a social graph with n nodes and m edges given as a stream of edges, when the graph follows a power law degree distribution, in the regime when $m = O(n. \log n)$. A subgraph S is very dense if it has $\Omega(|S|^2)$ edges. We uniformly sample the edges with a Reservoir of size $k = O(\sqrt{n}. \log n)$. Our detection algorithm checks whether the Reservoir has a giant component. We show that if the graph contains a very dense subgraph of size $\Omega(\sqrt{n})$, then the detection algorithm is almost surely correct. On the other hand, a random graph that follows a power law degree distribution almost surely has no large very dense subgraph, and the detection algorithm is almost surely correct. We define a new model of random graphs which follow a power law degree distribution and have large very dense subgraphs. We then show that on this class of random graphs we can reconstruct a good approximation of the very dense subgraph with high probability. We generalize these results to dynamic graphs defined by sliding windows in a stream of edges.

Keywords: dense subgraphs; clustering; streaming; probabilistic analysis; random graphs; approximation

1. Introduction

We study the efficient detection of large dense subgraphs in social graphs, given as a stream of edges, by looking at only a small fraction of the stream. Our viewpoint comes from two constraints. First, given the massive size of social networks, our algorithms cannot store the graph in memory: the space used must be sublinear. Second, the dynamic feature corresponds to the evolution of the social network and our algorithms have to take a single pass over the stream. Given those two constraints, what kind of structure can be detected algorithmically?

Large dense subgraphs. Social networks such as Twitter evolve dynamically, and dense subgraphs appear and disappear over time as interest in particular events grows and wanes. How can we detect large dense subgraphs efficiently? The classical density is the ratio $\rho = |E[S]|/|S|$, where $E[S]$ is the set of edges internal to S. In the case of a stream of edges, the approximation of dense subgraphs is well studied in Bhattacharya et al. (2015), Epasto et al. (2015), Esfandiari et al. (2015), McGregor et al. (2015) and an $\Omega(n)$ space lower bound is known Bahmani et al. (2012). Social graphs define a specific regime for which we propose a streaming algorithm which uses $O(\sqrt{n}. \log n)$ space. Our density objective is however different.

Definition 1.1. *The (γ, δ)-large very dense subgraph* problem, where $\gamma \leq 1$, takes as input a graph $G = (V, E)$ and decides whether there exists an induced subgraph $S \subseteq V$ such that $|S| > \delta\sqrt{n}$ and $|E[S]| > \gamma|S|(|S| - 1)/2$.

---

*A preliminary version was presented at FODS 2020 (Foundations of Data Science) Conference

A *very dense subgraph*, also called a γ-clique, is a subset S such that the condition $|E[S]| > \gamma|S|(|S| - 1)/2$ holds. If you consider two random nodes of S, they are connected by an edge with probability γ. Observe that γ-cliques have the following nice structure, which does not hold for the usual measure $|E(S)|/|S|$: if S is a γ-clique then for any $2 \leq i \leq |S|$, there exists a set of size i that is a γ-clique. In fact, a random subset of S has this property on average. The (γ, δ)-*large very dense subgraph* problem is NP-hard and hard to approximate (Hastad, 1996), as it contains the maximum clique problem as the special case when $\gamma = 1$. This leads us to use a new notion of approximation, adapted to a specific distribution of the inputs.

Degree distribution and the Configuration model. A scale-free network is a network whose degree distribution asymptotically follows a power law: the fraction of nodes with degree d is proportional to $d^{-O(1)}$ for d tending to infinity. Many real-world networks are thought to be scale-free. In this article, we focus on social network models with a fixed degree sequence following a power law d^{-2}, which is such that the total number of edges is $m = O(n. \log n)$. The configuration model takes a feasible degree sequence and generates a multigraph with that degree sequence. Let μ denote the distribution of simple graphs obtained with the configuration model applied to the power law degree distribution d^{-2}. (We defer the discussion of how to get a simple graph from the multigraph to Section 3.1). The configuration model is a standard model used already in sociology in 1938 in a directed version (Moreno & Jennings, 1938), and also for modeling the World Wide Web (Newman et al., 2001), public opinion formation (Watts & Dodds, 2007), etc. There are other generative models such as the Preferential Attachment model (Barabasi & Albert, 1999), the Copy model (Kleinberg et al., 1999), and many others, see Kumar et al. (2000), Albert & Barabási (2000), Aiello *et al.* (2000), Newman (2010) for example.

One-sided-stochastic approximation. We relax the definition of a randomized algorithm A, with a one-sided approximation, where for Yes instances we consider the worst case, but for No instances we only consider random inputs for μ. Indeed, a standard way to relax a decision problem would be to design an ε-tester that decides whether the input is a Yes instance or is ε-far from Yes instances, i.e. within edit distance at least εm from any Yes instance; but for our problem, no graph is ε-far from Yes instances, so the notion of ε-tester is not the right one here.

A *one-sided stochastic randomized* Reservoir algorithm A for a language L satisfies the following two conditions:

- For all $x \in L$, $Prob_{\Omega}[A(x) \text{ accepts}] \geq 1 - \varepsilon$
- If $x \notin L$ is drawn from μ, $Prob_{\mu \times \Omega}[A(x) \text{ rejects}] \geq 1 - \varepsilon$

where Ω is the set of possible Reservoirs (subsets of edges with prescribed cardinality), with a uniform distribution. We design a one-sided randomized streaming algorithm for the (γ, δ)-large very dense subgraph problem and believe that this approximation for a distribution of inputs can also be useful in some other contexts.

Static results. We study how to decide the presence of large very dense subgraphs with a Reservoir sampling (Vitter, 1985) and how to reconstruct such a very dense subgraph from the samples.[1] We study the existence of giant components in the Reservoir using random graph techniques, adapted to graphs with this degree distribution. Indeed, the Reservoir is a random uniform sample of a power law graph, whereas the Erdös-Renyi model is a random uniform sample of the complete graph. If the graph has a very dense subgraph S of size $\Omega(\sqrt{n})$, then a Reservoir of size $\Omega(\sqrt{n}. \log n)$ has a giant component (Theorem 4.1). In order to detect the existence of a very dense subgraph, our first Algorithm DETECT(γ, δ) simply checks whether there is a large enough connected component in the Reservoir. The analysis relies on a Theorem by Molloy and Reed on asymptotic sequences, but there are additional difficulties here, due to the fact that our sequences are not deterministic but random.

We then analyze the situation when the graph does not have a large very dense subgraph: we take the configuration model of random graphs which follow a power law degree distribution (Newman, 2010, see also next section). We observe that in this case there is no giant component in the Reservoir (Lemma 3.1) and that Algorithm DETECT(γ, δ) is correct (Theorem 4.2). In Corollary 5.2, we show an $\Omega(\sqrt{n})$ space lower bound.

Given a graph that has a very dense subgraph, how can we not only detect its existence but also reconstruct it? We propose a simple algorithm (Algorithm 2) that uses the 2-core of the largest connected component of the Reservoir. We define a variant of the configuration model, giving a preference to edges inside a certain subset of the vertices, so that the graph contains a very dense subgraph, and show that in this case Algorithm 2 reconstructs an approximation of the very dense subgraph (Theorem 6.1.)

Social networks dynamics and dynamic results. We consider "bursty" windows when the number of edges in each window varies. Uniform sampling in a window is the Reservoir sampling (Vitter, 1985) and it has been generalized to overlapping windows in Babcock et al. (2002), Braverman et al. (2009). We turn to the dynamic case with sliding windows. We show in Corollary 7.1 that we can detect the existence of a large very dense subgraph by generalizing Algorithm DETECT(γ, δ) (Algorithm 3). The configuration model can be generalized to dynamic random graphs (Section 7.1), and then none of the successive Reservoirs have a giant component, Corollary 7.2. In the concentrated case for some time interval Δ, the random graphs during that interval have a very dense subgraph S and Algorithm 4 approximates S, Corollary 7.3.

Plan of the paper. In Section 2, we review the large dense subgraphs, Reservoir sampling for dynamic graphs, and random graphs. In Section 3, we describe random graphs which follow a power law degree distribution with or without large very dense subgraph. In Section 4, we present the Algorithm DETECT(γ, δ) and its analysis on positive and random instances. In Section 5, we study the space lower bounds. In Section 6, we show how to reconstruct a good approximation of large very dense subgraph from the samples. In Section 7, we generalize the approach to dynamic graphs defined by sliding windows.

2. Preliminaries

Throughout the article, we ignore the rounding of our parameters to the nearest integer, when it only has an impact on second-order terms. For example, we look for a very dense subgraph with at least $\delta\sqrt{n}$ nodes. We approximate the existence of a very dense subgraph in the regime where $m = O(n. \log n)$, observed in social graphs.

2.1 Large dense subgraphs

There are several definitions of a cluster in a graph (Aggarwal & Wang, 2010) and our Definition 1.1 assumes a large γ-clique, as setting the value of the parameter γ to 1 corresponds to a clique for S. Studies of the Web graph have previously used large bipartite cliques as a defining marker of web communities (Kleinberg et al., 1999). This differs from a common definition of dense subgraph according to which the average degree within the subgraph, $\rho(S) = E(S)/|S|$, must be greater than some fixed threshold λ. A classical NP-hard (Khuller & Saha, 2009) optimization problem is to approximate $\rho^* = Max_S\{\rho(S)\}$ and to find possible witnesses S^* for ρ^*, in particular when the graph is given as a stream of edges (McGregor et al., 2015; Bhattacharya et al., 2015; Esfandiari et al., 2015). If the entire graph is known, there are several classical techniques to find large γ-clique subgraphs. In this context, it is hard to approximately detect subgraphs S with large value of $|E(S)|/|S|$: an $\Omega(n)$ space lower bound is known Bahmani et al. (2012) based on the multiparty Disjointness problem (Bar-Yossef et al., 2004; Chakrabarti et al., 2003). As we are interested in

large dense subgraphs, i.e. $|S| \geq \Omega(\sqrt{n})$, we observe in Corollary 5.1 that in this case the space lower bound is reduced to $\Omega(\sqrt{n})$.

For a large very dense subgraph, we show in Corollary 5.2 that the same space lower bound $\Omega(\sqrt{n})$ applies. For a comparison, our algorithm uses space $O(\sqrt{n}.\log n)$.

2.2 Sampling social graphs and dynamic graphs

Social graphs have a specific structure: a specific degree distribution (a power law), and some very dense subgraphs. In our framework, we do not store the entire graph as we sample the edges and will only approximate these very dense subgraphs. For example, imagine that the entire graph is a complete graph S. Each edge is selected in the Reservoir with probability $\frac{k}{m}$, so the Reservoir follows the Erdös-Renyi model $G(n, p)$ (see next section) where $n = |S|$ and $p = \frac{k}{m}$. It is well-known that there is a phase transition at $p = 1/n$, for the emergence of a giant component. We first show that for a γ-clique subgraph, a giant component emerges at $p = 1/\gamma.n$. We then study sufficient conditions to observe a giant component in the Reservoir.

Dynamic graphs and models of densification in social graphs have been studied in Leskovec et al. (2005). Dynamic algorithms approaches are presented in Demetrescu et al. (2010). We consider sliding windows defined by two parameters: the time length τ, and the time step $\lambda < \tau$ such that λ divides τ. In a stream of edges $e_1, e_2, ...e_i...$, each edge has a timestamp. Let $t_1 = \tau$ and $t_i = \tau + \lambda.(i - 1)$ for $i > 1$. Let G_i denote the graph defined by the edges whose timestamps are in the time interval $[t_i - \tau, t_i]$. The graphs G_{i+1} and G_i share many edges: old edges of G_i are removed and new edges are added to G_{i+1}. We write $G(t)$ for the sequence $G_1, G_2, ...G_i$ until we reach time t.

A Reservoir sampling (Vitter, 1985) reads a stream of edges $e_1, e_2, ...e_i, ...e_n$ and selects k edges uniformly ($k \leq n$), i.e. each edge is chosen with probability $\frac{k}{m}$. It first initializes the Reservoir to $e_1, e_2, ...e_k$. Then for each i such that $k < i \leq n$, it selects e_i with probability k/i. If e_i is selected, it replaces a random element of the Reservoir, selected with probability $1/k$. In a dynamic window w_i, we assume we can maintain k uniform samples in each Reservoir R_i, using techniques presented in Babcock et al. (2002), Braverman et al. (2009).

2.3 Random graphs

2.3.1 Erdös-Renyi graphs

The classical Erdös-Renyi model $G(n, p)$ (Erdös & Renyi, 1960), generates a random graph with n nodes and edges are taken independently with probability p where $0 < p < 1$. The degree distribution is close to a Gaussian centered on $n.p$. A *giant component* is a connected component with $\Theta(n)$ vertices. A classical study is to find sufficient conditions so that the random graph has a *giant component*. In the Erdös-Renyi model $G(n, p)$, it requires that $p > 1/n$. If $p = c/n$ with $c > 1$, we can be more precise on the size of the giant component C and the size of the 2-core(C).

Definition 2.1. *The* 2-core(C) *is obtained from C by removing nodes of degree 1, iteratively.*

Theorem 2.1 (**6** from Pittel & Wormald, 2005). *Let $c > 1$ be fixed. Consider the Erdös-Renyi model $G(n, p)$ with $p = c/n$. Let C denote the largest connected component and 2-core(C) its 2-core. There exists b, t such that $t.e^{-t} = c.e^{-c}$ and $b = 1 - t/c$ such that:*

$$lim_{n \to \infty} I\!\!E |C|/n = b$$

$$lim_{n \to \infty} I\!\!E |2\text{-core}(C)|/n = b \cdot (1 - t)$$

In addition, variables $|C|$ and $|2\text{-core}(C)|$ are both Gaussian in the limit.

We will use this result in Section 6. A similar result (Theorem 5.1) is given in Ding et al. (2014). When c is a large constant (tending slowly to infinity), t is close to 0, $\Theta(ce^{-c})$, and so $b = 1 - \Theta(e^{-c})$ tends to 1, therefore both C and 2-core(C) have size tending to n with high probability. We use the following result which generalizes the Erdös–Renyi condition to observe a giant component of an arbitrary non-empty graph G. Let d_u be the degree of a node u.

When we write that a property holds *almost surely*, we mean that the probability is at least $1 - o(1)$ as n goes to infinity. [2]

The existence of giant components on dense graph sequences has been studied in Bollobas et al. (2010) and the phase transition occurs when $p = 1/\lambda_n$ where λ_n is the largest eigenvalue of the adjacency matrix. A direct consequence which we will use in Section 4.1 is the following lemma:

Lemma 2.1. *Let $\varepsilon > 0$ be fixed. If S is a γ-clique and $p > \frac{1+\varepsilon}{\gamma.|S|}$ then the random subgraph $\widehat{S}$ obtained in $G(|S|, p)$, almost surely has a giant component.*

Proof. It is well known that $\lambda_n \geq \gamma.n$ on γ-dense graphs of size n, the average degree. Using Bollobas et al. (2010)'s result on the phase transition on a γ-clique graph S, we directly obtain:

$$\frac{1}{\lambda_n} \leq \frac{1}{\gamma.|S|} < \frac{1+\varepsilon}{\gamma.|S|}$$

We conclude that if $p > \frac{1+\varepsilon}{\gamma.|S|}$, the random subgraph $\widehat{S}$ almost surely has a giant component. □

2.3.2 Power-law graphs

Most of the social graphs have a degree distribution close to a power law (such as a Zipfian distribution distribution where $Prob[d = j] = c/j^2$). The preferential attachment or the configuration model (Newman, 2010) provide models where the degree distribution follows such a power law.

In the configuration model, the degree distribution can be an arbitrary distribution $\mathscr{D}$: given n nodes, we fix the number $d_1, d_2, ...d_{max}$ of nodes of degree 1, 2, ... where d_{max} is the maximum degree.

Definition 2.2. *Given n, let $c > 0$ be such that $\sum_{i \geq 1} \lfloor nc/i^2 \rfloor = n$ (c is approximately $6/\pi^2$). In the Zipf degree sequence, there are $d_i(n) = \lfloor nc/i^2 \rfloor$ nodes of degree i for each $i > 1$. The number of nodes of degree 1 is either $\lfloor nc \rfloor$ or $\lfloor nc \rfloor - 1$, chosen so that the sum of degrees is even; in the former case, all nodes have degree at least 1; in the latter case, there is one node of degree 0.*

Lemma 2.2. *For all n, the Zipf degree sequence is feasible: there exists a graph with that degree sequence.*

Proof. The definition of the degree sequence guarantees that the sum of degrees is even, and it is easy to see that the sequence then satisfies the Erdös-Gallai condition (Erdós & Gallai, 1960). □

Lemma 2.3. *The Zipf degree sequence satisfies the following elementary properties.*

- *The maximum degree is $dmax = \sqrt{cn}$.*
- *The total number m of edges satisfies*

$$\left| m - \frac{cn \ln (cn)}{4} \right| \leq \frac{cn}{2}$$

and the average vertex degree satisfies

$$\left| \mathbb{E}(d_u) - \frac{c \ln (cn)}{2} \right| \leq c$$

- *Let $d = \delta \sqrt{n} = dmax/\sqrt{a}$ with $a = c/\delta^2$. The number of nodes of degree greater than or equal to $\delta \sqrt{n}$ is*

$$\sqrt{cn} \left[\sum_1^{a-1} \frac{1}{\sqrt{i}} - \frac{a-1}{\sqrt{a}} \right] + O(1).$$

3. Graphs with and without large very dense subgraphs

In this section, we consider random graphs with a power-law degree distribution, first without large very dense subgraphs and then with such a subgraph.

3.1 Uniform configuration model

Consider an edge (u, v). If we cut it, it defines two half-edges also called *stubs*, attached to each node u and v.

Definition 3.1. *The* configuration model *generates a multigraph from a given degree sequence. The model constructs a uniform random perfect matching π between stubs as a symmetric permutation without fixed points: if $\pi(i) = j$, meaning that stub i maps to stub j, then $j \neq i$ and $\pi(j) = i$. To construct π, we greedily match stubs at random. We obtain a multigraph.*[3]

For a given degree sequence, let $\mathscr{D}$ denote the random variable equal to the degree of a uniform random node, and G denote the multigraph obtained from the degree sequence according to the configuration model.

Definition 3.2. *Let $d_i(n)$ denotes the number of vertices of degree i in an n-vertex graph. We define a well-behaved asymptotic degree distribution $(d_i(n))_{i,n}$:*

1. *It is feasible: for each n, there exists a graph with degree distribution $(d_i(n))_i$.*
2. *It is smooth: for each i, $\ell_i = \lim_{n \to \infty} d_i(n)/n$ exists.*
3. *The convergence of $i(i - 2)d_i(n)/n$ to its limit ℓ_i is uniform: $\forall \varepsilon \; \exists N \; \forall n > N \; \forall i \; |i(i-2) d_i(n)/n - \ell_i| < \varepsilon$.*
4. *$L(\mathscr{D}) = \lim_{n \to \infty} \sum_i i(i-2)d_i(n)/n$ exists and the convergence is uniform: $\forall \varepsilon \; \exists i_0 \; \exists N \; \forall n > N \; |\sum_{i \leq i_0} i(i-2)d_i(n)/n - L(\mathscr{D})| < \varepsilon$.*

Theorem 3.1. *(1(b) from Molloy–Reed, 1998) Consider an asymptotic degree sequence such that:*

1. *the asymptotic degree sequence is well-behaved,*
2. *$Q(\mathscr{D}) = \sum_i (i^2 - 2i)\ell_i$ is less than a constant less than 0,*
3. *the maximum vertex degree is less than $n^{1/9}$, and*
4. *the average vertex degree is $O(1)$.*

Then the following holds almost surely in the configuration model: the largest connected component of G has size at most $Bn^{1/4}$ for some constant B depending on $Q(\mathscr{D})$; no connected component of G has more than one cycle; and there are at most $2Bn^{1/4}$ cycles.

As noted in Molloy & Reed (1998), for a well-behaved sequence we have

$$L(\mathscr{D}) = \lim_{n \to \infty} \sum_i i(i-2)d_i(n)/n = \sum_i i(i-2) \lim_{n \to \infty} d_i(n)/n = Q(\mathscr{D})$$

We will apply Theorem 3.1 to the Reservoir R to bound the size of its largest connected component.

First we prove a simple property of the uniform configuration model.

Lemma 3.1. *In the uniform configuration model, G almost surely does not have a γ-clique of size $\Omega(\sqrt{n})$.*

Proof. Assume, for a contradiction, that there exists a set S of size $c\sqrt{n}$ which is γ-dense. Let T denote the set of vertices of S of degree less than $(\gamma/2)(|S|-1)/2$ and let $S_1 = S \setminus T$. $E(S_1)$ contains at least $(\gamma/2)|S|(|S|-1)/2$ edges, and all vertices of S_1 have degree at least $(\gamma/2)c\sqrt{n}$ in V. Let A denote the set of vertices of V of degree at least $(\gamma/2)c\sqrt{n}$ in V. Its edge set contains $E(S_1)$, so $|E(A)| \geq \gamma c^2 n/4$. Using the degree distribution, $|A| = \theta(\sqrt{n})$ and A is γ' dense. This event occurs with small probability as $m = \theta(n.\log n)$ edges. The probability that a random edge, created by the random stub-matching algorithm, is in $E(A)$ is $\theta(n/n.\log n) = \theta(1/\log n)$. The expected number of edges internal to A is at most $\theta(n/\log n)$ and the expected density $|E(A)|/|A|^2 = \theta(1/\log n)$, i.e. $o(1)$. By Markov's inequality, the probability that A is γ'-dense is $O(1/\log n)$, hence we obtain a contradiction. $\square$

3.2 Models with a large very dense subgraph

The configuration model generates a random graph which follows a power law degree distribution, as explained in Section 2.3.2, with a uniform matching between the stubs. With a power law degree distribution, the random graph G has a giant component almost surely. We show in Theorem 4.2 that the Reservoir almost surely does not have a giant component and that G does not have a large very dense subgraph.

We now define a concentrated model, a stochastic power law model that defines a graph that contains a large very dense subgraph. In Theorem 6.1, we will prove that in that case, there is an algorithm, Algorithm 2, that not only decides but also reconstructs (approximately) the underlying very dense subgraph.

Our model for a random graph with a very dense subgraph (concentrated model) is as follows. Let $\delta \leq \sqrt{c}/2$ and $0 < \gamma \leq 1$ be fixed. We now construct a graph with a γ-clique .

- We attach d stubs (half-edges) to each node v of degree d of the distribution $\mathscr{D}$,
- Let S be a set of $\delta\sqrt{n}$ nodes chosen arbitrarily among all nodes that have degree at least $\delta\sqrt{n}$.
- For each vertex u of S, we mark $\delta\sqrt{n} - 1$ stubs of u at random. For each pair of marked stubs between different vertices, we put in the graph the edge between those stubs independently and with probability γ, creating a random graph over S distributed as the Erdös-Renyi model $G(|S|, \gamma)$;
- The remaining unmarked stubs of V and the marked stubs not chosen in (1) are matched uniformly.

By Lemma 2.3, for $\delta \leq \sqrt{c}/2$ there are at least $\delta\sqrt{n}$ nodes of degree at least $\delta\sqrt{n}$, so there exist such subsets S. Let $0 < \gamma \leq 1$ be fixed. Notice that the second procedure may add some edges in S.

4. Existence of a large very dense subgraph

Let C be the largest connected component of the Reservoir of size $k = \Theta(\alpha\sqrt{n}\log n)$. In order to decide the graph property P: *is there is a γ-clique of size greater then $\delta.\sqrt{n}$?* Consider this simple algorithm, where $\alpha > 1/\gamma\delta$ is an auxiliary parameter.

Algorithm 1:

---

1 **Algorithm DETECT(γ, δ)**
2 **Input:** a stream of edges of a graph G.
3 **Output:** whether G contains a large very dense subgraph
4 Let $\alpha = \Theta(1/(\gamma\delta))$.
5 • Use Reservoir sampling to maintain a reservoir R of size $k = \Theta(c.\alpha\sqrt{n}\log n)/4)$.
6 •Let C denote the vertices of the largest connected component of R.
7 • Accept iff $|C| \geq \Theta(n^{1/8}\log^2 n)$

The bound $n^{1/8}\log^2 n$ is a direct application of the Molloy-Reed Theorem 3.1. The sampling rate $\frac{k}{m} = \frac{c.\alpha\sqrt{n}\log n/4}{cn.\log n/4} = \frac{\alpha}{\sqrt{n}}$. We show the correctness of this algorithm in two steps. In Section 4.2, we show in Theorem 4.2 that Algorithm DETECT(γ, δ) rejects almost surely in the uniform case. In the next section, we show in Theorem 4.1 that Algorithm DETECT(γ, δ) accepts almost surely if there is a large γ-clique .

4.1 Analysis when there is a large very dense subgraph

In this section, we analyze the size of the largest connected component C of the Reservoir used for the detection of a very dense subgraph (Algorithm DETECT(γ, δ)). The following theorem formalizes the fact that Algorithm DETECT(γ, δ) is correct with high probability on any graph that contains a large γ-clique and has $m = cn.\log n/4$ edges.

Theorem 4.1. *Assume that G contains a γ-clique on S where $|S| > \delta.\sqrt{n}$ and has $m = cn.\log n/4$ edges. If $\alpha > \frac{(1+\varepsilon)}{\gamma.\delta}$, then Algorithm DETECT$(\gamma, \delta)$ accepts almost surely.*

Proof. If S is a γ-clique and G has $m = cn.\log n/4$ edges:

$$\frac{k}{m} = \frac{\alpha}{\sqrt{n}} > \frac{(1+\varepsilon)}{\gamma.|S|}$$

By Lemma 2.1 there is a giant component as

$$\frac{k}{m} > \frac{(1+\varepsilon)}{\gamma.|S|}$$

Hence Algorithm DETECT(γ, δ) accepts almost surely. □

The above theorem shows that on positive instances, Algorithm DETECT(γ, δ) is almost surely correct. What about negative instances? We observe that there exists an input graph G that does not have a γ-clique of size strictly greater than $\varepsilon\sqrt{n}$, yet which Algorithm DETECT(γ, δ) (incorrectly) accepts. G consists of a clique K of size $\varepsilon\sqrt{n}$ and of a path of size $n - |K|$. With high probability, the Reservoir contains a component of size at least 90% of K, and will therefore accept, incorrectly.

However this input is somewhat pathological. In Theorem 4.2, we will prove that, assuming that G is drawn from the following stochastic power law model, the algorithm is correct on G with high probability.

4.2 Analysis in the uniform case

Theorem 4.2. *In the uniform configuration model (Subsection 2.3.2), Algorithm* DETECT(γ, δ) *Rejects almost surely.*

The Reservoir R is constructed by a three-step process, which we call the *Configuration-first* process:

1. Take the Zipf degree sequence (Definition 2.2).
2. Use the configuration model (Definition 3.1) to generate a multigraph G from the Zipf degree sequence
3. Take a random uniform sample (without replacement) of k edges from G to define the Reservoir R.

Instead, we will analyze the following process, which we call the *Configuration-last* process:

1. Take the Zipf degree distribution (Definition 2.2) as the degree sequence of the overall graph G.
2. Take a random uniform sample (without replacement) of $2k$ stubs, determining the degree sequence in R.
3. Use the Configuration model (Definition 3.1) to generate a random multigraph R from that degree sequence.

Lemma 4.1. *The above two processes give the same distribution for multigraph R.*

Proof. The configuration model is simply a greedy matching of stubs. $\qquad\square$

Proof. (of Theorem 4.2) By Lemma 4.1, we will analyze the second process for generating R. The plan is to apply Theorem 3.1 to the configuration model generated from the degree sequence of R. One difficulty is that that degree sequence is not deterministic but random. Theorem 3.1 assumes a deterministic degree sequence for each n. In our setting, we have a distribution of degree sequences for each n.

We analyze the properties of the random degree sequence in R, so as to prove that with probability $1 - o(1)$ it satisfies the assumptions of Theorem 3.1. We define one degree sequence for each n. We will prove that this asymptotic sequence satisfies the assumptions of Theorem 3.1.

The proof relies on a series of technical Lemmas, that are deferred to the next subsections. The degree distribution of the Reservoir is well-behaved: the first property (feasibility) holds by definition. The second property (smoothness) holds by Lemma 4.3 for $i = 1$ and by Lemma 4.4 for $i \geq 2$. The third property (uniform convergence) also holds by Lemma 4.6. The fourth property (uniform limit) holds by Lemma 4.8. Thus the degree sequence is well-behaved. As to the other assumptions of Theorem 3.1, the second one (Q negative) follows from the previous Lemmas, the third one (maximum degree) from Lemma 4.5, and the fourth one (average degree) from Lemma 4.7.

By Theorem 3.1, we then have that, with probability $1 - o(1)$, the largest connected component of R has size $O(n_R^{1/4} \log n_R)$. Since $n_R = O(\sqrt{n} \log n)$, this is $O(n^{1/8} \ln^{5/4} n)$, and then Algorithm DETECT(γ, δ) rejects. $\qquad\square$

4.2.1 Degree distribution in the Reservoir

Let $\mathscr{D}_R$ denote the distribution of degrees in R. Let N_R denote the number of nodes spanned by the edges of R, with expectation $n_R = \mathbb{E}[N_R]$, and let X_i be the random variable equal to the

number of nodes of degree i in R, with expectation $x_i = \mathbb{E}[X_i]$, where the expectation is over the Configuration-last process. We will use the following basic fact:

Lemma 4.2. *If $n \geq 0$ and $0 \leq x \leq 1$ then $(1-x)^n \geq 1 - nx$.*

Lemma 4.3. *In the Reservoir R:*

- $X_1 \leq N_R \leq 2k$.
- $\mathbb{E}(X_1) = x_1 \geq 2k\left(1 - O\left(\frac{\ln \ln n}{\ln n}\right)\right)$.
- *Let $\eta > 0$. Then, with probability at least $1 - \eta$ we have: $X_1 \geq 2k\left(1 - O\left(\frac{\ln \ln n}{\eta \cdot \ln n}\right)\right)$.*
- *The limit in probability of the ratio X_1/N_R exists and is equal to $\ell_1 = 1$: that is, $\mathrm{plim}_{n \to \infty} X_1/N_R = 1$.*

Proof. The first statement is obvious: the number of vertices of degree 1 is at most the toal number N_R of vertices spanned by the k edges of R, which is at most $2k$.

For the second statement, we start from an exact expression for the expected number of nodes of degree 1 in R. A vertex that has degree j in G has degree 1 in R if and only if the Reservoir picks exactly one of its j edges, which has probability $j \cdot (\alpha/\sqrt{n})(1 - \alpha/\sqrt{n})^{j-1}$. Since there are $\lfloor cn/j^2 \rfloor$ vertices of degree j in G, we can write:

$$x_1 = \sum_{j=1}^{\sqrt{c.n}} \left\lfloor \frac{c.n}{j^2} \right\rfloor \cdot j \frac{\alpha}{\sqrt{n}} \left(1 - \frac{\alpha}{\sqrt{n}}\right)^{j-1}.$$

Let $\varepsilon = \frac{2 \ln \ln n}{\alpha \sqrt{c} \ln(n)}$. We use Lemma 4.2 to write:

$$\left(1 - \frac{\alpha}{\sqrt{n}}\right)^{j-1} \geq \begin{cases} 1 - \varepsilon \alpha \sqrt{c} & \text{if } j \leq \varepsilon \sqrt{cn} \\ 0 & \text{otherwise.} \end{cases}$$

Thus

$$x_1 \geq (1 - \varepsilon \alpha \sqrt{c}) \sum_{j=1}^{\varepsilon \sqrt{c.n}} \left\lfloor \frac{c.n}{j^2} \right\rfloor \cdot j \frac{\alpha}{\sqrt{n}}.$$

On the other hand, since $2k$ equals the expected sum of the degrees of vertices in R, and a vertex of degree j in G has expected degree $j \cdot \alpha/\sqrt{n}$ in R, we can therefore write

$$2k = \sum_{j=1}^{\sqrt{c.n}} \left\lfloor \frac{c.n}{j^2} \right\rfloor \cdot j \frac{\alpha}{\sqrt{n}}$$

and

$$x_1 \geq (1 - \varepsilon \alpha \sqrt{c}) \left(2k - \sum_{j=\varepsilon \sqrt{cn}+1}^{\sqrt{c.n}} \left\lfloor \frac{c.n}{j^2} \right\rfloor \cdot j \frac{\alpha}{\sqrt{n}} \right).$$

Now we can bound the last term by

$$\sum_{j=\varepsilon \sqrt{cn}+1}^{\sqrt{c.n}} \left\lfloor \frac{c.n}{j^2} \right\rfloor \cdot j \frac{\alpha}{\sqrt{n}} \leq c\sqrt{n}\alpha \int_{\varepsilon \sqrt{cn}}^{\sqrt{c.n}} \frac{dt}{t} = c\sqrt{n}\alpha \ln(1/\varepsilon).$$

Recall that $k = m\alpha/\sqrt{n}$ and $m \sim cn \ln(n)/4$, so that $2k \sim c\sqrt{n}\alpha \ln(n)/2$. We obtain:

$$x_1 \geq 2k\left(1 - 2\frac{\ln(1/\varepsilon)}{\ln(n)}\right)(1 - \varepsilon\alpha\sqrt{c}).$$

Substituting the value of ε, we obtain $x_1 \geq 2k\left(1 - O\left(\frac{\ln\ln n}{\ln n}\right)\right)$.

For the third statement, we use Markov's inequality on $Y_1 = 2k - X_1$. We have: $Y_1 \geq 0$, and $\mathbb{E}(Y_1) = 2k - x_1 \leq 2kO\left(\frac{\ln\ln n}{\ln n}\right)$ by Lemma 4.3, so $\Pr(Y_1 > \mathbb{E}(Y_1)/\eta) < \eta$, hence the result.

For the fourth statement, observe that $X_1/N_R \leq 1$ always. Let $\eta = (\ln\ln n)^2/\ln n$. Then, by the previous statement, with probability $1 - \eta = 1 - o(1)$ we have

$$\frac{X_1}{N_R} \geq \frac{X_1}{2k} \geq 1 - O(\ln\ln n/\ln n)\frac{1}{\eta} = 1 - O(1/\ln\ln n), \text{ i.e.}$$

$$\Pr\left(|\frac{X_1}{N_R} - 1| \geq \varepsilon\right) = O\left(\frac{\ln\ln n}{\varepsilon\ln n}\right),$$

hence for all ε we have $\lim_{n\to\infty} Pr\left[|\frac{X_1}{N_R} - l_1| \geq \varepsilon\right] = 0$, therefore $plim_{n\to\infty}\frac{X_1}{N_R} = 1$. □

Lemma 4.4. *In the Reservoir R:*

- *for $i \geq 2$, the expected number of nodes of degree i satisfies $\mathbb{E}(X_i) = x_i \leq \frac{(\alpha\sqrt{c})^i}{i!.i-1}.\sqrt{c.n}$.*
- *The limit in probability of the ratio X_i/N_R exists and is equal to $\ell_i = 0$: that is, $plim_{n\to\infty}X_i/N_R = 0$.*

Proof. For the first statement, consider that a vertex that has degree j in G has degree i in R if and only if the Reservoir picks exactly i of its j edges; since there are $\lfloor c.n/j^2 \rfloor$ vertices of degree j in G, we can write:

$$x_i = \sum_{j=i}^{\sqrt{c.n}} \left\lfloor \frac{c.n}{j^2} \right\rfloor \cdot \binom{j}{i} \cdot \left(\frac{k}{m}\right)^i \cdot \left(1 - \frac{k}{m}\right)^{j-i}$$

Since $\left(1 - \frac{k}{m}\right)^{j-i} \leq 1$ and $\binom{j}{i} \leq j^i/i!$,

$$x_i \leq c.n.\left(\frac{k}{m}\right)^i \cdot \frac{1}{i!} \sum_{j=i}^{\sqrt{c.n}} \frac{1}{j^2}.j^i.$$

We use $\sum_{j=i}^{\sqrt{c.n}} j^{i-2} \leq \int_0^{\sqrt{cn}} t^{i-2}dt = \frac{1}{i-1}\sqrt{cn}^{i-1}$.

$$x_i \leq c.n.\left(\frac{\alpha}{\sqrt{n}}\right)^i \cdot \frac{1}{i!(i-1)}\sqrt{cn}^{i-1} = \frac{(\alpha\sqrt{c})^i}{i!(i-1)}\sqrt{cn},$$

hence the first statement. The second statement follows from $X_i \leq N_R - X_1$ and the fact that $plim_{n\to\infty}X_1/N_R = 1$. □

Lemma 4.5. *Let $\eta > 0$. Let i^* denote the maximum vertex degree in the Reservoir R. Then:*

- *With probability at least $1 - \eta$, we have $i^* \leq \log(n/\eta)/\log\log(n/\eta)$.*
- *$\mathbb{E}(i^*) = O(\log n/\log\log n)$.*

Proof. Consider the first statement. For any i we have $\Pr(i^* < i) = 1 - \Pr(\exists v \text{ of degree} \geq i)$. Let $i \geq 2\alpha\sqrt{c}$. With the union bound, we compute:

$$\Pr(\exists v \text{ of degree} \geq i) \leq \sum_{i \leq j' \leq \sqrt{cn}} \left\lfloor \frac{cn}{j^2} \right\rfloor \sum_{i \leq j \leq j'} \binom{j'}{j} \left(\frac{k}{m}\right)^j \left(1 - \frac{k}{m}\right)^{j'-j}.$$

A short calculation ensues: $\binom{j'}{j} \leq j'^j / j!$, and $1 - k/m \leq 1$. Exchanging the order of summation, the right hand side is bounded by

$$\sum_{j=i}^{\sqrt{cn}} \frac{cn(k/m)^j}{j!} \sum_{j'=j}^{\sqrt{cn}} j'^{j-2} \leq \sum_{j=i}^{\sqrt{cn}} \frac{cn(k/m)^j}{j!} \frac{1}{j-1} \sqrt{cn}^{j-1}.$$

Substituting $k/m = \alpha\sqrt{c}$, the right-hand side is bounded by

$$\sqrt{cn} \sum_{j=i}^{\sqrt{cn}} \frac{1}{j-1} \frac{(\alpha\sqrt{c})^j}{j!}.$$

Since $j \geq i \geq 2\alpha\sqrt{c}$, we have $\alpha\sqrt{c}/j \leq 1/2$, and so we can use bound the sum by the first term times $(1 + 1/2 + 1/4 + \cdots) \leq 2$:

$$\sqrt{cn} \sum_{j=i}^{\sqrt{cn}} \frac{1}{j-1} \frac{(\alpha\sqrt{c})^j}{j!} \leq 2\sqrt{cn} \frac{1}{i-1} \frac{(\alpha\sqrt{c})^i}{i}.$$

Fix $\eta > 0$. The right-hand side is less than η for $i = \frac{\lg(n/\eta)}{\lg\lg(n/\eta)}$.

Consider the second statement. We write

$$\mathbb{E}(i^*) = \sum_{i \geq 1} \Pr(i^* \geq i) \leq i_0 + \sum_{i \geq i_0} \Pr(\exists v \text{ of degree} \geq i).$$

For any $i_0 \geq 2\alpha\sqrt{c}$, again we can write

$$\mathbb{E}(i^*) \leq i_0 + \sum_{i \geq i_0} \sqrt{cn} \sum_{j=i}^{\sqrt{cn}} \frac{1}{j-1} \frac{(\alpha\sqrt{c})^j}{j!} \leq i_0 + 2 \sum_{i \geq i_0} \sqrt{cn} \frac{1}{i-1} \frac{(\alpha\sqrt{c})^i}{i!}$$

$$\leq i_0 + 4\sqrt{cn} \frac{1}{i_0 - 1} \frac{(\alpha\sqrt{c})^{i_0}}{i_0!}$$

Minimizing the right-hand side over $i_0 \geq 2\alpha\sqrt{c}$ gives $\mathbb{E}(i^*) = O(\lg n / \lg\lg n)$. □

Lemma 4.6 (Uniform convergence)**.**

$$plim_{n\to\infty} \sum_{i \geq 2} \frac{i(i-2)X_i}{N_R} = 0.$$

Proof. We must prove that for all $\varepsilon > 0$, we have $\Pr\left(\sum_{i \geq 2} i(i-2)X_i/N_R > \varepsilon\right) = o(1)$.

Let $\eta > 0$. We first use $N_R \geq X_1$ to infer $i(i-2)X_i/N_R \leq (i(i-2)X_i/X_1$.

- By Lemma 4.3, with probability at least $1 - \eta$ we have $X_1 \geq 2k(1 - (1/\eta)) \ln\ln n / \ln n$.
- By Lemma 4.5, with probability at least $1 - \eta$, the maximum degree i^* satisfies $i^* = O(\log n / \log\log n)$.

- Let $2 \leq i \leq i^*$. By Markov's inequality, with probability at least $1 - \eta$ we have: $\sum_{i \geq 2} i(i-2)X_i \leq \sum_{i \geq 2} i(i-2)x_i/\eta$. Using Lemma 4.4, this implies: $\sum_{i \geq 2} i(i-2)X_i \leq \sum_{i \geq 2} \frac{(\alpha\sqrt{c})^i}{(i-1)!} \cdot \sqrt{c.n} \frac{1}{\eta} \leq \alpha\sqrt{c} e^{\alpha\sqrt{c}} \sqrt{cn}/\eta$.

Combining, with probability $1 - O(\eta)$, all the above statements hold, implying:

$$\sum_{i \geq 2} i(i-2)\frac{X_i}{N_R} \leq \alpha\sqrt{c} e^{\alpha\sqrt{c}} \sqrt{cn} \frac{1}{\eta} \frac{1}{2k(1-(1/\eta))\ln\ln n/\ln n}$$

Since $\alpha, c = \Theta(1)$ and $k = \theta(\sqrt{n}\log n)$, the above equation means that with probability at least $1 - O(\eta)$:

$$\sum_{i \geq 2} i(i-2)\frac{X_i}{N_R} \leq \frac{O(1)}{\eta \log n - \log\log n}$$

Let $\eta = (1/\varepsilon + \log\log n)/\log n = o(1)$. Then

$$\Pr\left(\sum_{i \geq 2} i(i-2)X_i/N_R \leq O(\varepsilon)\right) = 1 - O(\eta) = 1 - o(1) \qquad \square$$

Lemma 4.7. *In the Reservoir R, with probability at least $1 - o(1)$, the average vertex degree is at most 2. Here we take expectations over the Configuration.last model, and then compute the average degree over all vertices in the Reservoir.*

Proof. The average vertex degree in R equals $2k/N_R$. The probability that it exceeds 2 equals the probability that $N_R < k$, which is at most the probability that $X_1 < k$, which by Lemma 4.3 is $O(\log\log n/\log n) = o(1)$. $\qquad \square$

We now consider the last condition for a well-behaved degree sequence: the limit $L(\mathscr{D}_R) = plim_{n \to \infty} \sum_i i(i-2)X_i/N_R$ exists, is equal to -1, and the convergence is uniform: $\forall \varepsilon \exists i_0 \exists N \forall n > N$ $|\sum_{i \leq i_0} i(i-2)X_i/N_R - L(\mathscr{D}_R)| < \varepsilon$. We will take $i_0 = 2$.

Lemma 4.8 (Uniform limit). *The limit $L(\mathscr{D}_R) = plim_{n \to \infty} \sum_i i(i-2)X_i/N_R$ exists and is equal to -1. Moreover, the convergence is uniform: Let $\eta > 0$. Then, for every $n > N$, with probability at least $1 - \eta$ we have:*

$$\left|\sum_{i=1}^{i=2} i(i-2)\frac{X_i}{N_R} - L(\mathscr{D}_R)\right| < \frac{O(\ln\ln n)}{\eta \log n} = o(1)$$

$$and \quad \left|\sum_{i \geq 3} i(i-2)\frac{X_i}{N_R}\right| < \frac{O(1)}{\eta \log n - \log\log n} = o(1)$$

Proof. This follows from Lemmas 4.3 and 4.6. $\qquad \square$

4.2.2 High probability analysis for sequences of degree distributions

We now consider a sequence of degrees indexed by both i and n, where for each n: we pick a random degree sequence in R obtained by the second step of the Configuration.last process. We use the following construction to couple the degree sequences as n varies.

Given n and a degree sequence in the Configuration.last process, let us write i^* for the maximum degree and $\left(X_i^{(n)}\right)$ for the degree sequence $\left(X_1^{(n)}, X_2^{(n)}, ...X_{i^*}^{(n)}\right)$. We define a function d whose

value is small with high probability and whose is value is 1 when two of the Molloy–Reed conditions are not satisfied: the maximum degree is greater then $N_R^{1/9}$ and the average degree $2k/N_R$ is greater than 2.

$$
d((X_i^{(n)})) = \begin{cases} 1 & \text{if } i^* > N_R^{1/9} \\ & \text{or } 2k/N_R > 2 \\ \max\left(\left|\dfrac{X_1^{(n)}}{N_R} - 1\right|, \left|\sum_{i \geq 2} i(i-2)\dfrac{X_i^{(n)}}{N_R}\right|\right) & \text{otherwise.} \end{cases}
$$

In Lemma 4.8, we separated the two expressions $\left|\dfrac{X_1^{(n)}}{N_R} - 1\right|$ and $\left|\sum_{i \geq 2} i(i-2)\dfrac{X_i^{(n)}}{N_R}\right|$ whose limits are small. For each n, we sort the degree sequences in increasing order according to d, producing an order on those degree sequences, $\left(X_i^{(n,1)}\right), \left(X_i^{(n,2)}\right), \ldots$.

To each number $0 < t < 1$, we associate the unique sequence $\left(X_i^{(n,r)}\right)$ such that:

$$
\sum_{j<r} \Pr\left(\left(X_i^{(n,j)}\right)\right) < t < \sum_{j \leq r} \Pr\left(\left(X_i^{(n,j)}\right)\right)
$$

Note that r is a function of t and n. This provides the desired coupling: for each t, as n spans the natural integers we obtain a collection of degree sequences, one for each n. We wish to apply the Molloy–Reed Theorem (Theorem 3.1) to the resulting sequence of degree sequences $\left(X_i^{(n,r)}\right)$ and conclude.

Proof of theorem 4.2. Consider a sequence indexed by both i and n, generated as above. We check that this asymptotic degree sequence is almost surely well-behaved and the other three assumptions of the Molloy–Reed Theorem hold.

First, it is well-behaved: it is feasible by construction. The smoothness condition, i.e. the existence of $plim_{n \to \infty} X_i^{(n,r)}/N_R$ is proved in the Lemma 4.3 for $i = 1$ and in the Lemma 4.4 for $i > 1$. The uniform convergence of $plim_{n \to \infty} \sum_{i \geq 2} \dfrac{i(i-2)X_i^{(n,r)}}{N_i}$ is proved in the Lemma 4.6. The existence and uniform convergence of $\sum_{i \geq 1} i(i-2)\dfrac{X_i^{(n,r)}}{N_R}$ is proved in the Lemma 4.8. The condition on the maximum degree, less than $N_R^{1/9}$, is proved in the Lemma 4.5 and the condition on the average degree in the Lemma 4.7. The coeffcient Q is -1 and we can therefore apply the Molloy–Reed Theorem 3.1. The conclusion, i.e. the bound on the size of the largest connected component C holds almost surely. There exists a constant B such that: $Pr_t[|C| \leq B.N_R \leq B.k^{1/4}] = 1$. Hence:

$$
Pr_{\text{Configuration.last}}[|C| \leq B.k^{1/4}] = Pr_{\mu.\Omega}[|C| \leq k^{1/4}] \to_{n \to \infty} 1
$$

The algorithm DETECT(γ, δ) tests if $|C| \leq k^{1/4} \leq n^{1/8} \log^2 n$, hence it is correct with h.p. $\qquad \square$

5. Space lower bounds

We reduce a hard problem for the one-way communication complexity to the existence of an $(1, \delta)$-large very dense subgraph and assume the reader familiar with the concepts of communication complexity (Kushilevitz & Nisan, 1997). The multiparty disjointness problem in the one-way

communication model is defined as follows. There are q players and for each $j = 1, ...q$, player j has an n-bit vector $x_j = x_{j,1} \ldots x_{j,n}$. In the restricted problem:

- either all vectors x_j are pairwise distinct (i.e. there is no j, j', i such that $x_{j,i} = x_{j',i} = 1$),
- or there exists a unique i^* such that $\bigwedge_j x_{j,i^*} = 1$, and all vectors x_j are otherwise pairwise distinct (i.e. there is no $j, j', i \neq i^*$ such that $x_{j,i} = x_{j',i} = 1$).

In the one-way communication model, information is only sent from a player j to a player j' such that $j' > j$, and the last player, player q, must decide whether there is an i such that $\bigwedge_j x_{j,i} = 1$.

Lemma 5.1 ((Chakrabarti et al., 2003)). *The restricted multiparty disjointness problem with one-way communication and q players requires communication complexity $\Omega(n/q)$.*

This lower bound is used to obtain other lower bounds for a range of problem. An $\Omega(n)$ lower bound is presented in Bahmani et al. (2012) for the α-approximation of the maximum density ratio ρ^*, i.e. to find a subgraph with a vertex S such that $\rho(S) \geq \rho^*/\alpha$.

Lemma 5.2 ((Bahmani et al., 2012)). *An $\sqrt{q/2}$-approximation streaming algorithm for the maximum density ratio requires space $\Omega(n/q)$.*

We keep the same reduction presented in Bahmani et al. (2012): it reduces the restricted multiparty disjointness problem with one-way communication and q players to the $\sqrt{q/2}$-approximation of the maximum density ratio ρ^*. Consider an instance of the restricted q-party disjointness problem with 1-way communication. Player j holds boolean variables $x_{j,1}, x_{j,2}, \cdots, x_{j,n}$, for each $j = 1, 2, \ldots, q$. We construct a graph G such that:

- G is the union of n disjoint graphs $G_1, ...G_n$, each over q vertices. For $i = 1, 2, \ldots, n$, the nodes of G_i are denoted $u_{1,i}, u_{2,i}, ...u_{q,i}$.
- If $x_{j,i} = 1$, we add the $q - 1$ edges from the node $u_{j,i}$ to all the other nodes $u_{j',i}$ of G_i, for $j' \neq j$.

For a Yes instance of the multiparty disjointness problem, i.e. $\bigwedge_j x_{j,i^*} = 1$, the graph G_{i^*} is a clique of size q and the maximum density ratio is therefore $\rho^* = (q - 1)/2$. For a No instance, G is a forest where each tree is of depth 1 and the maximum density ratio is $\rho^* = (q - 1)/q = 1 - 1/q$.

Let an input stream with the edges of player 1, then of player 2, ... then of player q coming last. A streaming algorithm for ρ^* with approximation less than $\sqrt{q/2}$ and space $o(n/q)$ could decide between a Yes and a No instance of the q-multiparty disjointness problem. If we assume that S must be large, let $q = \delta\sqrt{n}$ and we obtain:

Corollary 5.1. *A $n^{1/4}$-approximation streaming algorithm for the maximum density ratio for S of size at least $2\sqrt{n}$ requires space $\Omega(\sqrt{n})$.*

If we consider the very dense criterium, i.e. $\gamma = 1$, we obtain our lower bound:

Corollary 5.2. *The detection of a (γ, δ) large very dense subgraph requires $\Omega(\sqrt{n})$ space.*

Proof. Let $q = \delta\sqrt{n}$ in the reduction of Lemma 5.2 in Bahmani et al. (2012). A Yes (resp. No) instance of the multiparty disjointness problem is reduced to a Yes (resp. No) instance of the $(1, \delta)$-Large very dense subgraph problem. It implies the space lower bound of 5.1. $\square$

Algorithm 2:

---

1 Static very dense subgraph estimation.

2 • *Let C be the largest connected component of the reservoir R.*

3 • *If* $|C| < n^{1/8} \ln^2 n$ *then Reject, else output* 2-core(C).

---

6. Concentrated case: Reconstructing a complete subgraph

In this section, we propose an algorithm (Algorithm 2) to approximately reconstruct a very dense subgraph. We show that in the Concentrated model with a clique S (a γ-clique with $\gamma = 1$), the output approximates S with high probability.

To reconstruct a clique, based on Algorithm DETECT(γ, δ), it would be tempting to output the largest connected component in the Reservoir. However, in the Concentrated model with a clique S, such an algorithm would overestimate the size of S and output many vertices besides the vertices of S. Instead, we observe that many of the extraneous vertices appear to be leaves of the connected component of the Reservoir, and can thus be eliminated by outputting the 2-core instead of the full connected component. Formalizing this intuition leads to the following algorithm and result.

We now assume that the parameter α depends on n and slowly grows to infinity. For example $\alpha = O(\ln \ln \ln n)$ so that the analysis of the previous sections still hold. If α is constant, we just approximate S within a constant factor.

Theorem 6.1. *Assume the concentrated model (Section 3.2) with a clique S, and let $\widehat{S} = 2$-core(C) denote the output of Algorithm 2. If $1/\delta = o(\alpha)$ and $\alpha = O(\ln \ln \ln n/\sqrt{c})$, then $|S \setminus \widehat{S}|$ and $|\widehat{S} \setminus S|$ almost surely are both $o(|S|)$.*

Since S is a clique in this section, the edges of the reservoir R that are internal to S are exactly distributed according to $G(n, p)$. Notice that:

- If $v \in S$, then the degree of v is $O(\sqrt{n})$
- The degree of v in the Reservoir is $\log n$

Recall Theorem 2.1 which gives us an estimate on the size of $|C|$ ($|C| \simeq b.n$) and of its 2-core(C) ($|2\text{-core}(C)| \simeq b.(1 - t).n$). If S is a clique, we can use this result and conclude that the 2-core(C) is of size $O(n)$ and in S with high probability. The proof follows the successive steps, where results are taken with high probabilities:

- The Reservoir has no cycle disjoint of S, Lemma 6.1 ,
- The 2-core(C) consists of elements of S and possibly elements of $V − S$ which belong to cycles that go through S,
- Lemmas 6.2 and 6.3 show that very few nodes of $V − S$ belong to such cycles.

Lemma 6.1. *With high probability, in the restriction of the Reservoir to $V − S$, the total size of connected components that have a cycle is at most $O(n^{1/4})$.*

Proof. Use Molloy-Reed, Theorem 3.1. □

Nodes of S have a high degree. Some nodes in $V − S$ have also a high degree. Let us study false positives for S, that is, vertices u that belong to 2-core(C) but not to S.

$$\mu = \Pr\left[(u \in V − S \wedge u \in 2\text{-core}(C))\right]$$

One possibility is that u is connected to 2-core(C) $\cap$ S by two disjoint paths. We first consider the case where those paths have length 1.

Lemma 6.2. *Let Z denote the number of vertices u in $V - S$ incident to two edges of R into S, $(v_1, u) \in R$ and $(v_2, u) \in R$ with $v_1, v_2, \in S$. Then:*

$$\mathbb{E}[Z] \leq \sqrt{n}\frac{O(\alpha^2)}{(\log n)^2}$$

Proof. By linearity of expectation, the desired quantity is equal to:

$$A = \sum_{u \in V-S} \Pr\left[\text{there exist } v_1, v_2 \in S \text{ such that } (v_1, u), (u, v_2) \in R\right].$$

Edges (v_1, u) and (v_2, u) are border edges of S. By definition of S-concentrated dynamics and of the size of R, S has $O(\alpha\sqrt{n})$ border edges in R. By construction of the power law graph, each of those edges is attached to a random stub of $V - S$. Since there are $\sim cn\log n/2$ stubs in $V - S$, the probability that such an edge is attached to a stub of u is $\Theta(d_u/(n\log n))$. We will also use

$$\sum_{u \in V-S} d_u^2 \sim \sum_{i=1}^{i=\sqrt{cn}} c.n.i^2/i^2 = \sum_{i=1}^{i=\sqrt{cn}} c.n = c\sqrt{c}.n.\sqrt{n} = \Theta(n^{3/2}).$$

$$A = \sum_{u \in V-S} \binom{O(\alpha\sqrt{n})}{2}(\Theta(d_u/n.\log n))^2 = \frac{O(\alpha^2)}{n.(\log n)^2}\sum_{u \in V-S} d_u^2$$

$$= \frac{O(\alpha^2)\sqrt{n}}{(\log n)^2} \qquad \square$$

Lemma 6.3. *Given $i \geq 2$, let Y denote the number of chains $(v, u_1, u_2, \ldots, u_i, v') \in S \times (V \setminus S) \times \cdots \times (V \setminus S) \times S$ of R. Then:*

$$\mathbb{E}[Y] \leq \sqrt{n}\frac{O(\alpha^{i+1})}{(\log n)^{i+1}}.$$

Proof. We will give the full proof for $i = 2$, and the generalization to $i \geq 2$ is immediate. Fix two nodes $u_1, u_2 \in V - S$.

$$\Pr\left[(u_1, u_2) \in R\right] = \Pr\left[(u_1, u_2) \in G\right]\Pr\left[(u_1, u_2) \in R|(u_1, u_2) \in G\right]$$

$$\sim \frac{d_{u_1}.d_{u_2}}{m}.\frac{O(\alpha)}{\sqrt{n}}$$

For u_1, the probability to be connected to S is:

$$\Pr\left[\exists v \in S{:}(v, u_1) \in R\right] = \frac{d_{u_1}}{m}O(\alpha)\sqrt{n}.$$

Similarly for u_i,

$$\Pr\left[\exists v' \in S{:}(u_i, v') \in R\right] = \frac{d_{u_i}}{m}O(\alpha)\sqrt{n}.$$

Combining, the probability that there exists $v, v' \in$ with a path (v, u_1, u_2, v') in R is bounded by

$$\Pr\left(\exists v, v' \in S{:}(v, u_1, u_2, v') \text{ path in } R\right) = \frac{d_{u_1}^2 d_{u_2}^2}{m^3}O(\alpha^3)\sqrt{n}.$$

Algorithm 3:

1 **Dynamic very dense subgraph detection** (t):
2 • *let C_i be the largest connected component of the reservoir R_i of size $k = \Theta(\alpha \sqrt{n} \log n)$ at time $t_i \leq t$.*
3 • *If there is an i such that $|C_i| \geq n_i^{1/8} \ln^2 n_i$ Accept, else Reject.*

Summing over $u_1, u_2 \in V - S$ and recalling that $\sum_u d_u^2 = O(n^{3/2})$ and that $m = \Theta(n \log n)$, we obtain that the expected number of chains, for $i = 2$, is

$$\sqrt{n} \frac{O(\alpha^3)}{(\log n)^3}.$$

The generalization to $i > 2$ is straightforward. □

Proof of Theorem 6.1. We use Theorem 2.1 to prove that 2-core(C) contains a large fraction of the elements of S. Indeed, the theorem means that $|2\text{-core}(C) \cap S| \geq |S|b(1 - t)$ with $b = 1 - t/c_1$ and $te^{-t} = c_1 e^{-c_1}$, where the probability that an edge of S is put in the Reservoir is $c_1/|S|$. We have $|S| = \delta \sqrt{n}$ and $\alpha/\sqrt{n} = c_1/|S|$. Hence $c_1 = \alpha.\delta$.

Assume $\alpha >> 1/\delta$. Then $c_1 \to \infty$, so $t \to 0$ and $b \to 1$, and then $|S|b(1 - t) \sim |S|$, so $|S \setminus \widehat{S}| = o(|S|)$.

To analyze $|\widehat{S} \setminus S|$, observe that the nodes of $V - S$ that belong to 2-core(C) either have two disjoint paths leading to S, or belong to a cycle of $V - S$.

The nodes that belong to a cycle of $V - S$ are few in number, $O(n^{1/4})$ by Lemma 6.1.

Concerning the nodes that have two disjoint paths leading to S, Lemmas 6.2 and 6.3 give: The number of nodes u in $V - S$ that have two disjoint paths leading to S is at most:

$$\sum_{i \geq 1} i\sqrt{n} \frac{O(\alpha^{i+1})}{(\log n)^{i+1}} \simeq \sqrt{n} \frac{O(\alpha^2)}{(\log n)^2}.$$

Thus the expected total number of nodes of $V - S$ that belong to 2-core(C) is $o(|S|)$. □

7. Dynamic graphs

Consider the sequence of graphs G_i defined by the edges in each window w_i. We keep a Reservoir R_i but only store the 2-core of the large connected components. We extend our model of random graphs with or without a very dense subgraph to the dynamic case.

Let P be the graph property: there is a γ-clique of size greater than $\delta.\sqrt{n}$. How do we decide $\Diamond P(t)$, i.e. there is a window w_i at some time $t_i \leq t$ such that G_i has a γ-clique of size greater than $\delta.\sqrt{n}$? Recall that $\alpha = \Theta(1/\gamma.\delta)$ as in the static case.

Consider the following Dynamics applied to a given graph G: remove $q \geq 2$ random edges, uniformly on the set of edges of G, freeing $2.q$ stubs. In the case of the uniform Dynamics, we generate a new uniform matching on these free stubs to obtain G'.

In the case of the S-concentrated Dynamics, we have fixed some subset S of size $\delta.\sqrt{n}$ among the nodes of high degree and some γ. Consider the q edges that we remove. Partition then into internal edges $E(S)$, external edges in $E(\bar{S})$ and cut edges of $E(S, \bar{S})$. There are three cases and the analysis generalizes the static case.

7.1 Dynamic models

For the S-concentrated Dynamics, we have fixed some subset S of size $\delta.\sqrt{n}$ among the nodes of high degree and some γ. Partition the q removed edges into internal edges $E(S)$, external edges in $E(\bar{S})$ and cut edges of $E(S, \bar{S})$. There are three cases:

1. The graph G had a γ-clique subgraph and we will maintain such a very dense subgraph. We rematch the corresponding stubs of each class with a uniform matching. We conserve the same number of edges in $E(S)$, $E(\bar{S})$ and $E(S, \bar{S})$ and maintain a γ-clique subgraph.
2. The graph had no γ-clique subgraph. Let q' be the number of edges in $E(S, \bar{S})$. With $\lfloor q'/2 \rfloor$ new edges in S, we don't have a γ-clique subgraph. If q' is even, we match uniformly all the stubs in S. If q' is odd, we only take $q' - 1$ stubs, match uniformly all the stubs in S and leave one edge in $E(S, \bar{S})$.
3. The graph had no γ-clique subgraph but with $q'' < \lfloor q'/2 \rfloor$ edges, we reach a γ-clique subgraph. We take the q'' stubs in S, match uniformly all the stubs in S, and match the other edges in $E(S, \bar{S})$ uniformly.

How does the distribution of random graphs evolve in time? Consider the Markov chain M where nodes are the possible graphs and transition probabilities $M(i, j)$ are the probabilities to obtain G_j from G_i with the process of removing random q edges from G_i and recombining the $2q$ stubs with the uniform or the S-concentrated Dynamics.

Lemma 7.1. *For the uniform Dynamics, the stationary distribution of the Markov chain M is uniform. For the S-concentrated S-Dynamics, the stationary distribution of the Markov chain M is uniform among all the graphs with a γ-clique S.*

Proof. By definition of the Markov chain, for all i, $\sum_j M(i, j) = 1$. The transition from G_i to G_j can also occur backwards from G_j to G_i, and similarly for many G_k which lead to G_i. Therefore for all j, $\sum_i M(i, j) = 1$ as it sums all possible transitions starting from G_j. The matrix M is then doubly stochastic and ergodic. Hence the stationary distribution is uniform. Fo the S-concentrated dynamics, we reach a γ-clique subgraph with the right number of edges. Each transformation of a γ-clique subgraph into another one is reversible. Hence the matrix M is also doubly stochastic and the stationary distribution is uniform. □

7.2 Deciding properties of dynamic random graphs

A general Dynamics is a function which chooses at any given time, one of the two strategies: either a uniform Dynamics or an S-concentrated Dynamics for a fixed S. An example is the Step Dynamics: apply the uniform Dynamics first, then switch to the S-dynamics for a time period Δ, and switch back to the uniform Dynamics. Notice that during the uniform Dynamics phase, there are no large components in the Reservoir. For the step phase, we store some components which will approximate S.

A stream $G(t)$ has a large γ-clique subgraph if there is G_i and S_i such that $|S_i| > \delta.\sqrt{n}$ and $\alpha > \frac{(1+\varepsilon)}{\gamma.\delta}$ for a $t_i \leq t$. Assume *Prob*[Algorithm 1 accepts] $\geq 1 - \eta$ in Theorem 4.1. We now show that the detection error of Algorithm 3 decreases.

Corollary 7.1. *For a stream which has a large γ-clique (which satisfies the conditions of Theorem 4.1) during a time interval $\Delta \geq \tau$, Algorithm 3 is such that Prob[Algorithm 3 accepts] $\geq 1 - \eta^{\Delta/\tau}$.*

Proof. There are $\lfloor (\Delta - \tau)/\lambda \rfloor$ different windows but $\lfloor \Delta/\tau \rfloor$ independent windows, i.e. windows with no overlap. The samples are then independent and we can then amplify the success probability. □

For the random graphs generated by the uniform Dynamics, the situation is different and the error will increase. If we assume that: *Prob*[Algorithm 1 Rejects] $\geq 1 - \eta'$, the detection error of Algorithm 3 is given by:

Algorithm 4:

---

1 **Dynamic very dense subgraph estimation Algorithm 4 (t, ε).**

2 • *Let $\widehat{S} = \emptyset$. Consider the first independent $I = \log(\varepsilon)/\log(\rho)$ windows w_i when C_i, the largest connected component of the reservoir R_i at time $t_i < t$ is such that $|C_i| > n_i^{1/8} \ln^2 n_i$.*

3 • *Then $\widehat{S} = \widehat{S} \cup 2\text{-}core(C_i)$.*

Corollary 7.2. *For a stream of random graphs which follow the uniform Dynamics, Algorithm 3 is such that:*

$$Prob[\text{ Algorithm } 3 \text{ Rejects }] \geq 1 - \left(\frac{t - \tau}{\lambda}\right) . \eta'$$

Proof. There are $\left\lfloor \frac{t-\tau}{\lambda} \right\rfloor$ windows and the error probability is less than the sum of the errors for each window. □

7.3 Dynamic estimation of S

Consider random graphs generated by a step Dynamics, i.e. a strategy which maintains an S concentration during a time interval $\Delta \geq \tau$. In this case, we can improve the quality of the approximation of Theorem 6.1. Assume $|S \setminus \widehat{S}| \leq \rho.|S|$ for Algorithm 2, where $\rho \leq 1$. Let ε be an arbitrary tolerated error, and say that $\widehat{S}$ ε-approximates S if $|S \setminus \widehat{S}| \leq \varepsilon.|S|$

Corollary 7.3. *For a stream generated by the step Dynamics during a time interval $\Delta \geq I.\tau$, Algorithm 4 will ε-approximate S almost surely.*

Proof. There are at least I independent windows where we have a large connected component. For each element in S, there is a probability ρ not to be selected in $\widehat{S}$ for each window. For I independent windows, the probability not to be selected is $\rho^I = \varepsilon$, hence a point is selected with probability $1 - \varepsilon$. By Theorem 6.1 very few nodes of $V - S$ are in $\widehat{S}$. □

7.4 Implementation

An implementation of the method, in de Rougemont & Vimont (2018), considers streams of edges defined by Twitter graphs[4] associated with tags (for example #bitcoin or #cnn). With windows of length $\tau = 1$ hour and step $\lambda = 30$ mins, there were $m = 20.10^3$ edges per window and the size k of the Reservoir was 500, of the order of $\sqrt{m}$. The #bitcoin stream has a unique giant component C in the Reservoir with approximately 100 edges, i.e. a compression factor of approximately 2.10^2. In practice, a giant component in the Reservoir is the witness of a large γ-clique, even though the coefficient γ can be small. The analysis of the variations of the sizes of the giant components over time is one of the motivations for the dynamic random graphs introduced in this article. Classical methods (Aggarwal & Wang, 2010) to detect communities in social graphs use Modules and Spectral techniques, which require to store the entire graph.

Competing interests. None.

Notes

1 In this article, whenever we speak of a "dense subgraph", we mean a *large* very dense subgraph, where the minimum size to qualify as large is specified in the theorem statements.

2 In some other contexts, it is sometimes called *asymptotically almost surely*.

3 Note that the distribution of the multigraph thus obtained is not uniform in general. We can obtain simple graphs, i.e. without self-loops or multi-edges by rejection sampling, and they satisfy the same properties. Since, for each simple graphs with that degree distribution, the number of executions leading to that simple graph is the same, the restriction of the distribution to simple graphs yielded by rejection sampling is uniform. (Molloy & Reed, 1998).

4 The nodes of a Twitter graph are the tags, either @*x* or #*t*. A tweet sent by @*x* which contains the tags @*y* and #*t* generates the edges (@*x*, @*y*) and (@*x*, #*t*). Given a tag or a set of tags, Twitter sends in a stream all the tweets which contain one of the selected tags. The stream of tweets is transformed into a stream of edges.

References

Aggarwal, C. C., & Wang, H. (2010). *Managing and mining graph data* (1st ed.). Springer Publishing Company, Incorporated.

Aiello, W., Chung, F., & Lu, L. (2000). A random graph model for power law graphs. *Experimental Mathematics*, *10*, 53–66.

Albert, R., & Barabási, A.-L. (2000). Topology of evolving networks: Local events and universality. *Physical Review Letters*, *85*, 5234–5237.

Babcock, B., Datar, M., & Motwani, R. (2002). Sampling from a moving window over streaming data. In *Proceedings of the thirteenth annual ACM-SIAM symposium on discrete algorithms* (pp. 633–634). SODA'02.

Bahmani, B., Kumar, R., & Vassilvitskii, S. (2012). Densest subgraph in streaming and mapreduce. *Proceedings of the VLDB Endowment*, *5*(5), 454–465.

Barabasi, A., & Albert, R. (1999). The emergence of scaling in random networks. *Science*, *286*, 509–512.

Bar-Yossef, Z., Jayram, T. S., Kumar, R., & Sivakumar, D. (2004). An information statistics approach to data stream and communication complexity. *Journal of Computer and System Sciences*, *68*(4), 702–732.

Bhattacharya, S., Henzinger, M., Nanongkai, D., & Tsourakakis, C. E. (2015). Space- and time-efficient algorithm for maintaining dense subgraphs on one-pass dynamic streams. Corr, abs/1504.02268.

Bollobas, B., Borgs, C., Chayes, J., & Riordan, O. (2010). Percolation on dense graph sequences. *The Annals of Probability*, *38*(1), 150–183.

Braverman, V., Ostrovsky, R., & Zaniolo, C. (2009). Optimal sampling from sliding windows. In *Proceedings of the twenty-eighth ACM SIGMOD-SIGACT-SIGART symposium on principles of database systems* (pp. 147–156). PODS'09.

Chakrabarti, A., Khot, S., & Sun, X. (2003). Near-optimal lower bounds on the multi-party communication complexity of set disjointness. In *IEEE conference on computational complexity* (pp. 107–117).

de Rougemont, M., & Vimont, G. (2018). The content correlation of streaming edges. In *IEEE international conference on big data* (pp. 1101–1106).

Demetrescu, C., Eppstein, D., Galil, Z., & Italiano, G. F. (2010). Dynamic graph algorithms. In M. J. Atallah, & M. Blanton, M. (Eds.), *Algorithms and theory of computation handbook*.

Ding, J., Lubetzky, E., & Peres, Y. (2014). Anatomy of the giant component: The strictly supercritical regime. *European Journal of Combinatorics*, *35*, 155–168.

Epasto, A., Lattanzi, S., & Sozio, M. (2015). Efficient densest subgraph computation in evolving graphs. In *Proceedings of the 24th international conference on world wide web* (pp. 300–310). WWW'15.

Erdós, P., & Gallai, T. (1960). Gráfok elöírt fokszámú pontokkal. *Matematikai lapok*, *11*, 264–274.

Erdös, P., & Renyi, A. (1960). On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, *5*, 17–61.

Esfandiari, H., Hajiaghayi, M., & Woodruff, D. P. (2015). Applications of uniform sampling: Densest subgraph and beyond. Corr, abs/1506.04505.

Hastad, J. (1996). Clique is hard to approximate within $n^{1-\varepsilon}$. In *Proceedings of the 37th annual symposium on foundations of computer science* (p. 627). FOCS'96. IEEE Computer Society.

Khuller, S., & Saha, B. (2009). On finding dense subgraphs. In *Proceedings of the 36th international colloquium on automata, languages and programming: Part I* (pp. 597–608). ICALP'09.

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. S. (1999). The web as a graph: Measurements, models, and methods. In T. Asano, H. Imai, D. T. Lee, S.-i. Nakano, & T. Tokuyama (Eds.), *Computing and combinatorics* (pp. 1–17). Berlin, Heidelberg: Springer.

Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., & Upfal, E. (2000). Stochastic models for the web graph. In *Proceedings of the 41st annual symposium on foundations of computer science* (p. 57). FOCS'00. IEEE Computer Society.

Kushilevitz, E, & Nisan, N. (1997). *Communication complexity*. Cambridge University Press.

Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining* (pp. 177–187). KDD'05.

McGregor, A., Tench, D., Vorotnikova, S., & Vu, H. T. (2015). Densest subgraph in dynamic graph streams. Corr, abs/1506.04417.

Molloy, M., & Reed, B. (1998). The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probability and Computing*, *7*(3), 295–305.

Moreno, J. L., & Jennings, H. H. (1938). Statistics of social configurations. *Jstor*, *1*(3/4), 342–374.

Newman, M. (2010). *Networks: An introduction*. Oxford University Press, Inc.

Newman, M. E. J., Strogatz, S. H., & Watts, D. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E, Statistical, Nonlinear, and Soft Matter Physics*, *64*(09), 026118.

Pittel, B., & Wormald, N. C. (2005). Counting connected graphs inside-out. *Journal of Combinatorial Theory, Series B*, *93*(2), 127–172.

Vitter, J. S. (1985). Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, *11*(1), 37–57.

Watts, D. J., & Dodds, P. S. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research*, *34*(4), 441–458.