Modeling the Small-World Phenomenon with Local Network Flow

Reid Andersen
University of California, San Diego

Fan Chung *
University of California, San Diego

Linyuan Lu
University of South Carolina

Abstract

The small-world phenomenon includes both small average distance and the clustering effect. Randomly generated graphs with a power law degree distribution are widely used to model large real-world networks, but while these graphs have small average distance, they generally do not exhibit the clustering effect. We introduce an improved hybrid model which combines a global graph (a random power law graph) with a local graph (a graph with high local connectivity defined by network flow). We present an efficient algorithm which extracts a local graph from a given realistic network. We show that the underlying local graph is robust in the sense that when our extraction algorithm is applied to a hybrid graph, it recovers the original local graph with a small error. The proof involves a probabilistic analysis of the growth of neighborhoods in the hybrid graph model.

1 Introduction

The small-world phenomenon usually refers to two distinct properties — *small average distance* and the *clustering effect* where two nodes with a common neighbor are more likely to be adjacent. These properties are ubiquitous in realistic networks. To model networks with the small-world phenomenon, it is natural to utilize randomly generated graphs with a power law degree distribution, where the fraction of nodes with degree k is proportional to $k^{-\beta}$ for some positive exponent β. This is based on the observations by several research groups that numerous networks, including Internet graphs, call graphs and social networks, have a *power law* degree distribution [1, 2, 3, 5, 6, 9, 10, 14, 15, 20, 22, 23, 24, 28, 26].

A random power law graph has small average distances and small diameter. It was shown in [11] that a random power law graph with a certain range of parameters almost surely has average distance of order $\log\log n$ and has diameter of order $\log n$. In contrast, the clustering effect in realistic networks is often determined by local connectivity and is not amenable to modeling using random graphs.

Most existing models that capture the clustering effect make random modifications to some underlying graph. Watts and Strogatz [29] introduced a model with an underlying graph consisting of vertices on the circle connected to their k nearest neighbors. Kleinberg [21] introduced a model for which the underlying graph is a grid. In both of these models, the graphs generated do not have a power law degree distribution, and each vertex has the same expected degree. The strict requirement that the underlying graph be a cycle or grid is unsatisfactory for modeling webgraphs or biological networks.

In this paper we introduce a hybrid graph model where the underlying graph can be any graph that satisfies a certain local connectivity property. This underlying local graph is then modified by adding the edges of a random power law graph, which we refer to as the global graph. The graphs generated by the

*Research supported in part by NSF Grants DMS 0457215, ITR 0205061 and ITR 0426858
[1]An extended abstract appeared in *Proceedings of the Third Workshop on Algorithms and Models for the Web-Graph*, 2004.

hybrid model have a power law degree distribution, small average distances between vertices, and allow very general underlying graphs.

The main difference between our hybrid model and the model introduced previously in [13] is that our notion of local connectivity is based on length-bounded network flows instead of length-bounded disjoint paths. Maximum length-bounded network flows can be computed efficiently using techniques for general fractional packing problems. This allows us to take a given real-world network and extract from it a highly connected local graph. We introduce such an extraction algorithm, and prove that when applied to graphs from the hybrid model, the extraction algorithm recovers the original local graph with only a small error. The extraction algorithm provides a way to partition a network into a local graph providing robust local connections, and a global graph providing small distances. Such a partition may have applications for clustering, routing, and graph visualization.

The paper is organized as follows. In section 2 we define random power law graphs. In section 3 we introduce our notion of local flow connectivity and define local graphs and hybrid graphs. In section 4 we present the `Extract` algorithm for extracting local subgraphs. In section 5 we state the main theorem: the `Extract` algorithm approximately recovers the original local graph when applied to a graph from the hybrid model. The proofs of the main theorem and a few related theorems are presented in sections 6 and 7. These proofs require bounds on neighborhood growth in the hybrid model, which are obtained in section 8 through a probabilistic analysis. In section 9 we point out that `Extract` can also be viewed as a clustering algorithm with some desirable properties, and we present drawings and examples.

2 Preliminaries

2.1 Notation

All graphs considered in this paper are undirected. Given a graph G we let $d_G(u, v)$ denote the graph distance between vertices u and v in G. We will use the following notation for vertex neighborhoods:

$$N_k^G(u) = \{v \in G \,|\, d_G(u, v) \leq k\},$$
$$\Gamma_k^G(u) = \{v \in G \,|\, d_G(u, v) = k\}.$$

When the graph is understood we will write $N_k(u)$ and $\Gamma_k(u)$. Given two sets of vertices A and B in G, we let $e_G(A, B)$ denote the number of edges in G with one endpoint in A and the other in B.

2.2 Random Graphs with Specified Expected Degrees

A random graph $G(\mathbf{w})$ with specified expected degree sequence $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ is formed by including each edge $v_i v_j$ independently with probability $p_{ij} = w_i w_j \rho$, where $\rho = (\sum w_i)^{-1}$. It is easy to check that vertex v_i has expected degree w_i. We assume that $\max_i w_i^2 < \sum_k w_k$ so that $p_{ij} \leq 1$ for all i and j. This condition also implies that the sequence w_i is graphical if the w_i are integers [16]. This model has a non-zero probability of self-loops, but the expected number of loops is much smaller than the total number of edges. The typical random graph $G(n, p)$ on n vertices with edge probability p is a special case of the $G(\mathbf{w})$ model where $\mathbf{w} = (pn, pn, \ldots, pn)$. For a subset S of vertices, we define

$$\mathrm{Vol}(S) = \sum_{v_i \in S} w_i \qquad \text{and} \qquad \mathrm{Vol}_k(S) = \sum_{v_i \in S} w_i^k.$$

We let d denote the average expected degree $Vol(G)/n$, and let $\tilde{d}$ denote the second-order average expected degree $\mathrm{Vol}_2(G)/\mathrm{Vol}(G)$. We also let m denote the maximum weight among the w_i.

The main results of this paper are stated for random graphs from the $G(\mathbf{w})$ model in terms of the parameters d, $\tilde{d}$, and m. However, we will mostly be interested in the special case where $G(\mathbf{w})$ is a random power law graph. The expected degree sequences of these graphs and the resulting values of the parameters d, $\tilde{d}$ and m are described in the next section.

2.3 Random Power Law Graphs

A random power law graph $M(n, \beta, d, m)$ is a random graph $G(\mathbf{w})$ whose expected degree sequence $\mathbf{w}$ is determined by the following four parameters.

- n is the number of vertices.
- $\beta > 2$ is the power law exponent.
- d is the average expected degree.
- m is the maximum expected degree and $m^2 = o(nd)$.

We let the i-th vertex v_i have expected degree

$$w_i = ci^{-\frac{1}{\beta-1}}$$

for $i_0 \leq i \leq i_0 + n$, for some c and i_0 (to be chosen later). It is easy to compute that the number of vertices of expected degree between k and $k+1$ is of order $c'k^{-\beta}$ where $c' = c^{\beta-1}(\beta-1)$, as required by the power law. To determine c, we consider

$$\begin{aligned} \mathrm{Vol}(G) &= \sum_i w_i = \sum_{i \geq i_0} ci^{\frac{1}{\beta-1}} \\ &\approx c\left(\frac{\beta-1}{\beta-2}\right) n^{1-\frac{1}{\beta-1}} \end{aligned}$$

Here we assume $\beta > 2$. Since $nd \approx \mathrm{Vol}(G)$, we choose

$$c = \left(\frac{\beta-2}{\beta-1}\right) dn^{\frac{1}{\beta-1}} \tag{1}$$

$$i_0 = n\left(\frac{d(\beta-2)}{m(\beta-1)}\right)^{\beta-1} \tag{2}$$

Values of $\tilde{d}$ for random power law graphs are given below (see [11]).

$$\tilde{d} = \begin{cases} (1+o(1))d\frac{(\beta-2)^2}{(\beta-1)(\beta-3)} & \text{if } \beta > 3. \\ (1+o(1))\frac{1}{2}d\ln\frac{2m}{d} & \text{if } \beta = 3. \\ (1+o(1))d^{\beta-2}\frac{(\beta-2)^{\beta-1}m^{3-\beta}}{(\beta-1)^{\beta-2}(3-\beta)} & \text{if } 2 < \beta < 3. \end{cases} \tag{3}$$

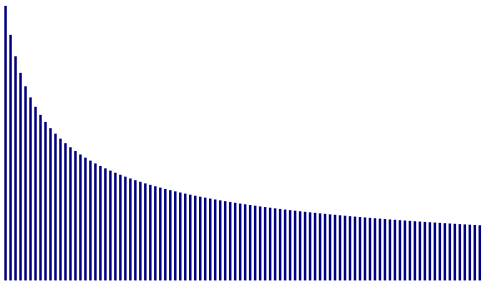

Figure 1: *Weight distribution $f(x)$.*

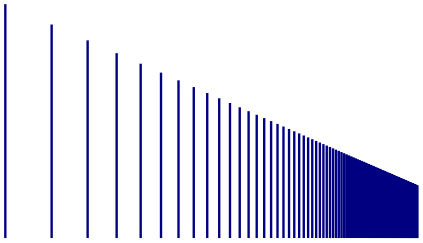

Figure 2: *Log-scale of figure 1.*

3 Local Graphs and Hybrid Graphs

3.1 Length-bounded network flow

There are a number of ways to define local connectivity between two given vertices u and v. One natural approach is to consider the connectivity through paths whose length is at most some fixed constant ℓ. We call such paths *short*, and let $a_\ell(u,v)$ be the maximum number of short edge-disjoint paths between u and v. Similarly, we let $c_\ell(u,v)$ be the minimum size of a short cut— a set of edges whose removal leaves no short path between the vertices. Both $a_\ell(u,v)$ and $c_\ell(u,v)$ are difficult to compute. Computing the maximum number of short disjoint paths is $\mathcal{NP}$-hard if $\ell \geq 4$ and $\mathcal{APX}$-hard if $\ell \geq 5$ [19]. Similar hardness results are known for computing the minimum size short cut [4]. The analogous version of the Menger's theorem for length restricted paths and cuts does not hold, and in fact $a_\ell(u,v)$ and $c_\ell(u,v)$ can be different by a factor of at least $\frac{\ell}{3}$ (see [25], [8], [4]). However we still have the trivial relations $a_\ell(u,v) \leq c_\ell(u,v) \leq \ell \cdot a_\ell(u,v)$.

Since the measures of local connectivity considered above are hard to compute, we will consider instead the maximum short flow between u and v. A *short flow* is a positive linear combination of short paths where no edge carries more than 1 unit of flow. Finding $f_\ell(u,v)$, the maximum size of a short flow between u and v, can be viewed as the linear programming relaxation of the maximum short disjoint paths problem. If we let A be the incidence matrix where each column represents a short path from u to v and each row represents an edge in the graph, then

$$f_\ell(u,v) = \max_{\mathbf{x}}\{\, \vec{\mathbf{1}}^T\mathbf{x} \mid A\mathbf{x} \leq \vec{\mathbf{1}}, \mathbf{x} \geq \vec{\mathbf{0}}\,\}. \tag{4}$$

The linear programming dual of the maximum short flow problem is a fractional cut problem. A short fractional cut is a weight function $w : E \to R^+$ such that $\sum_{e\in P} c(e) \geq 1$ for every short $u-v$ path P. The dual of the short maximum flow problem is the problem of finding a short fractional cut that minimizes $\sum_{e\in G} c(e)$. We let $w_\ell(u,v)$ denote the size of a minimum short fractional cut, and note that LP duality implies

$$a_\ell(u,v) \leq f_\ell(u,v) = w_\ell(u,v) \leq c_\ell(u,v). \tag{5}$$

Since all the coefficients in the incidence matrix, cost vector, and constraint vector in the linear program (4) are nonnegative, the maximum short flow problem belongs to a class of linear programs called fractional packing problems that can be solved efficiently by multiplicative update techniques (see for example [27], [30], [18]). In particular, it is easy to adapt the fractional packing algorithm of Garg and Könemann [18] to approximate to maximum short flow within a multiplicative factor of $(1+\epsilon)$ in time polynomial in ϵ and the size of the graph.

3.2 Local Graphs

We will use the maximum short flow $f_\ell(u,v)$ to define a measure of local connectivity and to define local graphs. Our definitions involve two parameters f and ℓ.

Definition 1 (Local Connectivity) *We say that vertices u and v are (f,ℓ)-connected if $f_\ell(u,v) \geq f$.*

Definition 2 (Local Graphs) *A graph L is an (f,ℓ)-local graph if for each edge $e = (u,v)$ in L, the vertices u and v are (f,ℓ)-connected in $L \setminus \{e\}$.*

We also define the notion of a local subgraph of a larger graph.

Definition 3 (Local Subgraphs) *A subgraph $L \subseteq G$ (not necessarily induced) is an (f,ℓ)-local subgraph of G if for each edge $e = (u,v)$ in L, the vertices u and v are (f,ℓ)-connected in $G \setminus \{e\}$.*

For a given graph G, we define $L_{f,\ell}(G)$ to be the set of edges $e = (u,v)$ where $f_\ell(u,v) \geq f$ in $G \setminus \{e\}$. It is clear than $L_{f,\ell}(G)$ is the unique largest local subgraph of G.

We also wish to consider the largest subgraph L of G that is an (f,ℓ)-local graph. Let $\hat{L}_{f,\ell}(G)$ be the union of all subgraphs of G that are (f,ℓ)-local. By definition, the union of two (f,ℓ)-local graphs is an (f,ℓ)-local graph, and so $\hat{L}_{f,\ell}(G)$ is the unique largest (f,ℓ)-local subgraph in G.

Changing the parameters f and ℓ yields different classes of local graphs. When one of the parameters is fixed we have the monotonicity results $L_{f,i} \subseteq L_{f,j}$ if $j \geq i$, and $L_{i,\ell} \subseteq L_{j,\ell}$ if $i \geq j$. We remark that $L_{f,\ell}(G)$ is not necessarily connected, and so the connected components of the local subgraph $L_{f,\ell}$ induce a partition $\Pi_{f,\ell}$ of the vertex set of G. In this case, the monotonicity results imply that $\Pi_{f,i}$ is a refinement of $\Pi_{f,j}$ if $j \geq i$, and $\Pi_{i,\ell}$ is a refinement of $\Pi_{j,\ell}$ if $i \geq j$.

3.3 Hybrid Power Law Graphs

A hybrid graph H is the union of the edge sets of an (f,ℓ)-local graph L and a random global graph $R = G(\mathbf{w})$ on the same vertex set. When generating the random graph R we allow the weights w_i from $\mathbf{w}$ to be assigned arbitrarily to the vertices of the local graph. Since the proofs will apply to any assignment of the weights to the vertices, we will ignore the particular assignment and simply write $H = L \cup R$.

We are most interested in the case where the global graph R is a power law graph $M(n,\beta,d,m)$. In this case, the hybrid graph will have small diameter and average distances, due to the following results on random power law graphs which appeared in [11].

Theorem 1 *For a random power law graph $R = M(n,\beta,d,m)$ and $\beta > 3$, almost surely, the average distance is $(1+o(1))\frac{\log n}{\log d}$ and the diameter is $O(\log n)$.*

Theorem 2 *For a random power law graph $R = M(n,\beta,d,m)$ and $2 < \beta < 3$, almost surely, the average distance is $O(\log\log n)$ and the diameter is $O(\log n)$. For a random power law graph $R = M(n,\beta,d,m,L)$ and $\beta = 3$, almost surely, the average distance is $O(\log n/\log\log n)$ and the diameter is $O(\log n)$.*

The diameter of the hybrid graph can be smaller than that of the random power law graph R if the local graph satisfies additional conditions. A local graph L is said to have isoperimetric dimension δ if for every

vertex v in L and every integer $k < (\log\log n)^{1/\delta}$, there are at least k^δ vertices in L of distance at most k from v. For example, the grid graph in the plane has isoperimetric dimension 2, and the d-dimensional grid graph has isoperimetric dimension d. The following results appeared in [13].

Theorem 3 *In a hybrid graph $H = R \cup L$ with $R = M(n, \beta, d, m, L)$ and $2 < \beta < 3$, suppose that the local graph has isoperimetric dimension δ, where $\delta \geq \log\log n/(\log\log\log n)$. Then almost surely, the diameter is $O(\log\log n)$.*

Theorem 4 *In a hybrid graph $H = R \cup L$ with $R = M(n, \beta, d, m, L)$ and $2 < \beta < 3$, suppose that the local graph has isoperimetric dimension δ. Then almost surely, the diameter is $O((\log n)^{1/\delta})$.*

Theorem 5 *In a hybrid graph $H = R \cup L$ with $R = M(n, \beta, d, m, L)$ and $2 < \beta < 3$, suppose that each vertex is within distance $\log\log n$ of some vertex of degree $\log n$. Then almost surely, the diameter is $O(\log\log n)$.*

4 Extracting the Local Graph

The following simple procedure `Extract` computes the largest (f, ℓ)-local subgraph of a given graph G.

`Extract`(f, ℓ): Given a graph G and parameters (f, ℓ), for each edge $e = (u, v)$ compute $f_\ell(u, v)$ in $G \setminus \{e\}$. Let L be the subgraph of G containing the edges $e = (u, v)$ for which $f_\ell(u, v) \geq f$.

It is easy to see that `Extract` computes $L_{f,\ell}(G)$. There is also a simple greedy algorithm to compute $\hat{L}_(f, \ell)(G)$, the largest subgraph L of G that is an (f, ℓ)-local graph.

`Recursive Extract`(f, ℓ): Given a graph G and parameters (f, ℓ), let $H_1 = G$. At iteration t of the algorithm, scan through the edges $e = (u, v)$ of H_t in any order and compute $f_\ell(u, v)$ in $H_t \setminus \{e\}$. Let $H_{t+1} \subseteq H_t$ be the set of edges in H_t where $f_\ell(u, v) \geq f$. Repeat until no edges are removed during an iteration of the algorithm, then output the remaining graph H_t.

Theorem 6 *For any graph G and any (f, ℓ),* `Recursive Extract`(f, ℓ) *returns $\hat{L}_{f,\ell}(G)$, the unique largest subgraph of G that is an (f, ℓ)-local graph.*

Proof: Given a graph G, let L be the graph output by `Recursive Extract`, A simple induction argument shows that each edge removed by the algorithm is not part of any (f, l)-local graph that is a subgraph of G, and thus $\hat{L}_{f,\ell}(G) \subseteq L$. Since no edges were removed from L in the final iteration of the algorithm, L is (f, l)-local and so $L \subseteq \hat{L}_{f,\ell}(G)$. Thus $L = \hat{L}_{f,\ell}(G)$. □

The algorithm `Extract` requires m maximum short flow computations, and `Recursive Extract` requires $O(m^2)$ maximum short flow computations.

4.1 Experiment

We have implemented the `Extract` algorithm and applied it to various hybrid graphs. For some hybrid graphs, the local graphs produced by `Extract` are almost perfectly recovered. In the next section we will present a theorem that makes this precise.

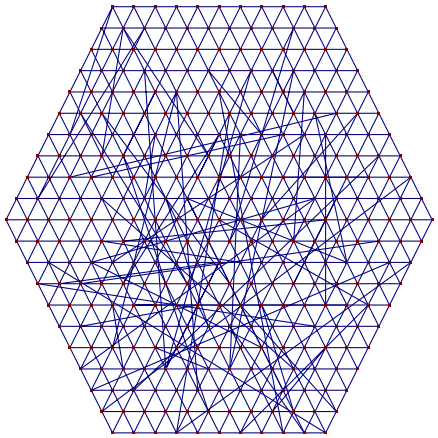

Figure 3: *A hybrid graph where the local graph is a hexagonal grid.*

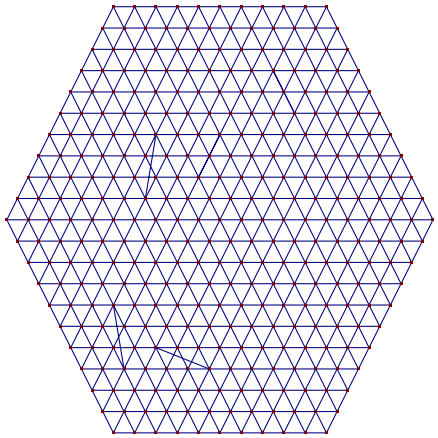

Figure 4: *After applying* `Extract` *with parameters* $f = 2.5$ *and* $l = 4$*, the local graph is almost perfectly recovered.*

5 Recovering the Local Graph

We now consider the problem of recovering a good approximation to the local graph L given a hybrid graph $H = L \cup R$. If we apply `Extract` with parameters (f, ℓ) to a hybrid graph $H = R \cup L$ where L is an (f, ℓ)-local graph, the algorithm will output $L_{f,\ell}(H)$. By definition we have $L \subseteq L_{f,\ell}(H)$, but $L_{f,\ell}(H)$ may also contain edges from $R \setminus L$. We now state our main theorem— that $L_{f,\ell}(H)$ is a good approximation of L if the random part R of the hybrid graph is sufficiently sparse.

Theorem 7 (Recovery Theorem) *Let $H = L \cup R$ be a hybrid graph where L is (f, ℓ)-local with bounded maximum degree M, and where $R = G(\mathbf{w})$ is a random graph with average expected degree d, second order average expected degree $\tilde{d}$, and maximum weight m. Let $L' = L_{f,\ell}(H)$. Let $\alpha > 0$ be some constant such that $\hat{d} = n^{\alpha}$ is an upper bound for $\tilde{d}$.*

$$\text{if} \quad \hat{d} \leq \left(\frac{nd}{m^2}\right)^{1/\ell} n^{-3/f\ell},$$

Then with probability $1 - O(n^{-1})$:

1. *The expected number of edges in $L' \setminus L$ is $O(\tilde{d})$.*
2. *$d_{L'}(x, y) \geq \frac{1}{\ell} d_L(x, y)$ for every pair of vertices $x, y \in L$.*

The proof of this theorem is contained in section 6. The statement of the theorem may become more clear by examining the corresponding result for the special case where all weights are equal and $G(\mathbf{w}) \sim G(n, p)$. In this special case the recovery theorem has a cleaner statement (Theorem 8). Also, we can show that the recovery theorem is tight in this special case in the sense that if d is slightly larger than $n^{\frac{1}{\ell}}$ we will have $\hat{L}_{f,\ell}(H) = H$, implying that neither `Extract` nor `Recursive Extract` recovers a good approximation to the original local graph.

Theorem 8 *Let $H = L \cup R$ be a hybrid graph as in Theorem 7 with $R = G(n,p)$ and $p = dn^{-1}$. If*

$$d \leq n^{\alpha} \leq n^{1/\ell} n^{-3/f\ell} \text{ for some constant } \alpha > 0,$$

Then with probability $1 - O(n^{-1})$, results (1)-(2) from Theorem 7 hold.

Theorem 9 *Let $H = L \cup R$ be a hybrid graph as in Theorem 7 with $R = G(n,p)$ and $p = dn^{-1}$, and let*

$$d \geq 6fn^{\frac{1}{\ell}} (\log n)^{\frac{1}{\ell}}.$$

With probability $1 - O(n^{-2})$, $\hat{L}_{f,\ell}(H) = H$.

The proofs of these results are contained in section 7. These results indicate that the term $\left(\frac{nd}{m^2}\right)^{1/\ell}$ in the Recovery Theorem is nearly the best possible, although we will not make this precise. When we deal with the $G(\mathbf{w})$ model, $\tilde{d}$ appears in place of d since we expect the volume of a small neighborhood to expand by a factor of $\tilde{d}$ as we increase the radius by one step.

6 Recovery Analysis

In this section we will prove the recovery theorem. We first introduce some notation. We say an edge in H is *global* if it is in $R \setminus L$. A global edge $e = (u,v)$ is *long* if $d_L(u,v) > \ell$ and *short* otherwise. We say a global edge *survives* if it is in $L_{f,\ell}(H) \setminus L$. In the propositions below we will show that there are not very many short global edges, and that long global edge are unlikely to survive. The proofs of these propositions are provided later in this section. Proving that there are few short global edges is easy, while the result for long edges requires a more detailed analysis that makes up the bulk of the remaining paper. This result also requires bounds on the growth of neighborhoods in hybrid graphs, which are presented in section 8. The recovery theorem follows easily from these propositions.

Proposition 1 (Short Edges) *The expected number of short edges in $L_{f,\ell}(H) \setminus L$ is $O(\tilde{d})$.*

Proposition 2 (Long Edges) *If the hypotheses from the recovery theorem hold, the probability that a given long edge survives is $O(n^{-3})$.*

Proof of recovery theorem: Since there are at most n^2 edges in R, combining Proposition 2 with the trivial union bound implies that with probability $1 - O(n^{-1})$ no long edges survive. In that case, $L' \setminus L$ contains only short edges, and there are $O(\tilde{d})$ of these by Proposition 1, so part (1) follows. To prove (2), note that if no long edges survive, then all edges in L' are short. If (u,v) is an edge in L', $d_L(u,v) \leq \ell$. If p' is a path between two vertices x, y in L' with length k, then by replacing each edge with a short path we obtain a path p in L between x and y with length at most ℓk. The result follows. □

6.1 A bound for sums

The following simple bound will be used several times is proving propositions 1 and 2.

Lemma 1 *Let X be some finite set with nonnegative weights $w(x)$, and let $A \subseteq X^k$ be a set of ordered k-tuples from X. If each element $x \in X$ appears in at most M elements of A, then*

$$\sum_{(x_{i_1} \ldots x_{i_k}) \in A} w(x_{i_1}) \cdots w(x_{i_k}) \leq M \sum_{x \in X} w(x)^k$$

Proof: Order the elements of X as $x_1 \ldots x_n$ such that $w(x_1) \geq \cdots \geq w(x_n)$. Let A_j be the collection of tuples $v \in A$ where j is the smallest index of any element in v. We have $|A_j| \leq M$, and $\cup_{j \in [1,n]} A_j = A$, so

$$\begin{aligned} \sum_{(x_{i_1} \ldots x_{i_k}) \in A} w(x_{i_1}) \cdots w(x_{i_k}) &\leq \sum_{j \in [1,n]} \sum_{(x_{i_1} \ldots x_{i_k}) \in A_j} w(x_{i_1}) \cdots w(x_{i_k}) \\ &\leq \sum_{j \in [1,n]} M w(x_j)^k \\ &\leq M \sum_{x \in X} w(x)^k \end{aligned}$$

□

6.2 Short Edges

If (u, v) is a short edge in R, it is possible that there is a short flow of size f from u to v in L. This means we can not say a short edge is unlikely to survive without placing additional assumptions on the local graph. However, an easy computation shows there are not likely to be many short edges in R.

Proof of Proposition 1: Let X be the number of short edges in R.

$$\begin{aligned} E[X] &= \sum_{\{(x,y) \,|\, x \in L,\, y \in N_\ell^L(x)\}} \Pr\left[\, (x,y) \in R \,\right] \\ &= \sum_{\{(x,y) \,|\, x \in L,\, y \in N_\ell^L(x)\}} w_x w_y \rho \end{aligned}$$

Since each vertex x appears in at most $2M^\ell$ terms in the above sum, we apply Lemma 1 to obtain

$$\begin{aligned} E[X] &\leq 2M^\ell \sum_{x \in G} w_x^2 \rho \\ &= 2M^\ell \tilde{d} \\ &= O(\tilde{d}) \end{aligned}$$

The proposition follows since $L_{f,\ell}(H) \setminus L$ is contained in R. □

6.3 Long Edges

Proving that a long edge (u, v) is unlikely to survive is similar to proving that few paths of length ℓ exists in H between u to v. If we were only dealing with a random graph R from the $G(n, p)$ model instead of a hybrid graph, one way to show this would be to bound the total number of vertices in the neighborhood $N_{\ell-1}(u)$ in the graph $R \setminus v$, and then to show that there are not likely to be f edges in R between that neighborhood and v. From this approach one can prove that paths of length ℓ between two given vertices are not likely to

appear until the expected degree of each vertex reaches roughly $n^{\frac{1}{\ell}}$, and at that point many short disjoint paths will appear with only a small increase in the average degree. This is reflected in Theorems 10 and 11.

To prove Proposition 2 we extend this line of reasoning to hybrid graphs. To deal with complications introduced by the local graph, we define modified neighborhoods $\bar{N}_k(u)$ and $\bar{\Gamma}_k(u)$ in the hybrid graph that avoid the local neighborhoods of v. In analogy to the case for $G(n,p)$ random graphs, we show that after $\ell - 1$ steps the total weight of vertices in these neighborhoods is not very large. We identify a set $S(u,v)$ of possible edges between the modified neighborhoods of u and local neighborhoods of v, and show that for an edge to survive at least f pairs among the set $S(u,v)$ must appear as edges in R. We also show that we can reveal the modified neighborhoods without examining any of the vertex pairs in $S(u,v)$. This implies that, after determining the modified neighborhoods, a pair $(x,y) \in S(u,v)$ appears as an edge in R with probability $w_x w_y \rho$. Since we bound the total weight of vertices in the modified neighborhoods, we can bound the expected number of edges among the pairs $S(u,v)$ that are included in R. If $\tilde{d}$ obeys the hypotheses of the recovery theorem, this expectation is small enough to imply that the edge (u,v) is unlikely to survive.

Definition 4 *$\bar{N}_k(u)$ and $\bar{\Gamma}_k(u)$*

For $k \in [0,\ell]$, let $\bar{N}_k(u)$ be the set of vertices y such that there exists a path $p = x_0 \dots x_k$ in H with $x_0 = u$ and $x_k = y$ obeying the following condition:

$$d_L(x_i, v) > \ell - i \text{ for all } i \in [0,k].$$

We define $\bar{\Gamma}_k(u)$ to be the corresponding strict neighborhood,

$$\bar{\Gamma}_k(u) = \left\{ y \mid y \in \bar{N}_k(u), y \notin \bar{N}_0(u) \cup \dots \cup \bar{N}_{k-1}(u) \right\}.$$

The following recursive definition of $\Gamma_k(u)$ will be useful, and is easily seen to be equivalent to the original.

$$\begin{aligned} \bar{\Gamma}_0(u) &= \{u\} \\ \bar{\Gamma}_k(u) &= \left\{ y \mid \begin{array}{l} y \notin N^L_{\ell-k}(v), \\ y \notin \bar{\Gamma}_0(u) \dots \bar{\Gamma}_{k-1}(u), \\ (x,y) \in H \text{ for some } x \in \bar{\Gamma}_{k-1}(u) \end{array} \right\} \end{aligned}$$

Definition 5 *$S(u,v)$ and $C(u,v)$*

We define $S(u,v)$ to be the set of vertex pairs

$$\bigcup_{k \in [1,\ell]} \left(\bar{\Gamma}_{k-1}(u) \times N^L_{\ell-k}(v) \right).$$

We define $C(u,v)$ to be the set of edges in H contained in $S(u,v)$.

Remark 1 *All the edges in C(u,v) are global edges.*

Proof: If $(x,y) \in \left(\bar{\Gamma}_{k-1}(u) \times N^L_{\ell-k}(v) \right)$, then $d_L(x,v) > \ell - (k-1)$ and $d_L(y,v) \le \ell - k$. Thus $d_L(x,y) \ge 2$, so (x,y) cannot be a local edge and must be global. □

Lemma 2 *If (u,v) is a surviving long edge then $|C(u,v)| \ge f$.*

Proof: We first show that every short path between u and v in H contains an edge from $C(u,v)$. Let $p = x_0 \dots x_k$ be a path of length $k \leq \ell$ between u and v in H. The last vertex on the path is $x_k = v$, so we have $d_L(x_k, v) = 0 \leq \ell - k$, and thus $x_k \notin \bar{N}_k(u)$. The first vertex on the path is $x_0 = u$, and thus $x_0 \in \bar{N}_0(u)$. Let $j \geq 1$ be the smallest integer such that $x_j \notin \bar{N}_j(u)$. By definition, $x_{j-1} \in \bar{N}_{j-1}(u)$, while $x_j \notin \bar{N}_j(u)$. This implies $d_L(x_j, v) \leq \ell - j$, so $x_j \in N^L_{\ell-j}(v)$. We now have that $x_{j-1}x_j$ is an edge in $\bar{N}_{j-1}(u) \times N^L_{\ell-j}(v)$. We conclude that $x_{j-1}x_j$ is an edge in $C(u,v)$ since

$$\bar{N}_{j-1}(u) \times N^L_{\ell-j}(v) \subseteq \bigcup_{k \in [1,j]} \left(\bar{\Gamma}_{k-1}(u) \times N^L_{\ell-k}(v) \right) \subseteq S(u,v).$$

We have shown that the set of edges $C(u,v)$ forms a short cut. Thus, we have

$$a_\ell(u,v) \leq f_\ell(u,v) \leq c_\ell(u,v) \leq |C(u,v)|.$$

Since (u,v) is a surviving edge, $f \leq f_\ell(u,v) \leq |C(u,v)|$. This completes the proof of the lemma. Notice that the lemma (and consequently the recovery theorem) still hold if we consider short disjoint paths $a_\ell(u,v)$ or short cuts $c_\ell(u,v)$ as our measure of local connectivity. □

Lemma 3 *If we condition on the values of the sets $\bar{\Gamma}_0(u) \dots \bar{\Gamma}_{\ell-1}(u)$, and let S be any set of edges contained in $S(u,v)$, then*

$$\Pr\left[\bigwedge_{(x,y) \in S} (x,y) \in R \mid \bar{\Gamma}_0(u) \dots \bar{\Gamma}_{\ell-1}(u) \right] = \prod_{(x,y) \in S} w_x w_y \rho \,.$$

Proof: Let $R_{(x,y)}$ denote the event that the vertex pair (x,y) is an edge in R. We are considering ordered pairs (x,y), but note that $R_{(x,y)}$ and $A_{(y,x)}$ are not independent since we are dealing with undirected graphs. We will determine $\bar{\Gamma}_0(u) \dots \bar{\Gamma}_{\ell-1}(u)$ sequentially by observing the events $R_{(x,y)}$ for a certain subset of vertex pairs $Q \subseteq V \times V$. From the recursive definition of $\bar{\Gamma}_k(u)$, it is clear that we can determine $\bar{\Gamma}_k(u)$ given $\bar{\Gamma}_{k-1}(u)$ by observing only the set of vertex pairs Q_k, where

$$Q_k = \bar{\Gamma}_{k-1} \times \left(V \setminus (N^L_{\ell-k} \cup \bar{\Gamma}_0(u) \cup \dots \cup \bar{\Gamma}_{k-1}(u)) \right).$$

It follows that Q_k does not contain any pairs from

$$\bigcup_{j \in [1,k]} \left(\bar{\Gamma}_{k-1}(u) \times N^L_{\ell-k}(v) \right) \quad \text{or} \quad \bigcup_{j \in [1,k]} \left(N^L_{\ell-k}(v) \times \bar{\Gamma}_{k-1}(u) \right).$$

We can determine $\bar{\Gamma}_0(u) \dots \bar{\Gamma}_{\ell-1}(u)$ by observing the pairs $Q = Q_1 \cup \dots \cup Q_{\ell-1}$. Since the neighborhoods $\bar{\Gamma}_j(u)$ are disjoint, we see that Q does not contain any pairs from

$$\bigcup_{k \in [1,\ell]} \left(\bar{\Gamma}_{k-1}(u) \times N^L_{\ell-k}(v) \right) \quad \text{or} \quad \bigcup_{k \in [1,\ell]} \left(N^L_{\ell-k}(v) \times \bar{\Gamma}_{k-1}(u) \right).$$

Thus, the set of vertex pairs Q for which we have observed $R_{(x,y)}$ in determining $\bar{\Gamma}_0(u) \dots \bar{\Gamma}_{\ell-1}(u)$ is disjoint from $S(u,v)$, and also disjoint from the set of pairs in $S(u,v)$ with the order of the vertices reversed. Thus the events $R_{(x,y)}$ with $(x,y) \in S(u,v)$ are independent of the events $R_{(x,y)}$ with $(x,y) \in Q$, and the claim follows. □

A bound on the total weight of pairs in $S(u,v)$ is given in the following lemma. The proof requires an analysis of the growth of the volumes of the neighborhoods $\bar{N}_k(u)$, is contained in section 8. Using the previous lemmas and the following lemma we can complete the proof of Proposition 2. .

Lemma 4 *With probability* $1 - e^{-\Omega(n^\alpha)}$,

$$\sum_{k\in[1,\ell]} Vol(\bar{\Gamma}_{k-1}(u))Vol(N^L_{\ell-k}(v)) \leq 4m^2(4M\hat{d})^{\ell-1}.$$

Proof of Proposition 2: If (u,v) is a surviving long edge, then $|C(u,v)| \geq f$ by Lemma 2. Let S^f denote the set of ordered f-tuples containing f distinct pairs from $S(u,v)$. Let B be the event that

$$\sum_{k\in[1,\ell]} Vol(\bar{\Gamma}_{k-1}(u))Vol(N^L_{\ell-k}(v)) \leq 4m^2(4M\hat{d})^{\ell-1}, \tag{6}$$

which occurs with probability $1 - e^{-\Omega(n^\alpha)}$ by Lemma 4. By the law of conditional probabilities,

$$\begin{aligned}\Pr[\,|C(u,v)| \geq f\,] &\leq \Pr[\,|C(u,v)| \geq f \mid B\,] + \Pr\left[\,|C(u,v)| \geq f \mid \bar{B}\,\right] \\ &\leq \Pr[\,|C(u,v)| \geq f \mid B\,] + e^{-\Omega(n^\alpha)}.\end{aligned}$$

To bound $\Pr[\,|C(u,v)| \geq f \mid B\,]$ we first determine $S(u,v)$ by conditioning on the sets $\bar{\Gamma}_0(u)\ldots\bar{\Gamma}_{\ell-1}(u)$. We can then write

$$\Pr\left[\,|C(u,v)| \geq f \mid \bar{\Gamma}_0(u)\ldots\bar{\Gamma}_{\ell-1}(u)\,\right] \leq \sum_{((x_1,y_1)\ldots(x_f,y_f))\in S^f} \Pr\left[\bigwedge_{i\in[1,f]} (x_i,y_i) \in R\right],$$

and apply Lemma 3 to obtain

$$\begin{aligned}\Pr\left[\,|C(u,v)| \geq f \mid \bar{\Gamma}_0(u)\ldots\bar{\Gamma}_{\ell-1}(u)\,\right] &\leq \sum_{((x_1,y_1)\ldots(x_f,y_f))\in S^f} \prod_{i\in[1,f]} w_{x_i}w_{y_i}\rho \\ &\leq \rho^f\left(\sum_{k\in[1,\ell]} Vol(\bar{\Gamma}_{k-1}(u))Vol(N^L_{\ell-k}(v))\right)^f.\end{aligned}$$

We then substitute the bound implied by event B in equation (6) to obtain

$$\begin{aligned}\Pr[\,|C(u,v)| \geq f \mid B\,] &\leq \rho^f\left(4m^2(4M\hat{d})^{\ell-1}\right)^f \\ &= \left(4m^2(4M\hat{d})^{\ell-1}\rho\right)^f.\end{aligned}$$

Since $\hat{d} \leq (\frac{nd}{m^2})^{1/\ell}n^{-3/f\ell}$ by the hypotheses of the recovery theorem,

$$\begin{aligned}\Pr[\,|C(u,v)| \geq f \mid B] \leq &\leq \left((4M)^\ell n^{-3/f}\right)^f \\ &= O(n^{-3}).\end{aligned}$$

Thus the probability that a given long edge survives is at most

$$O(n^{-3}) + e^{-\Omega(n^\alpha)} = O(n^{-3}).$$

□

7 Proof of Companion Theorems

We will use the following lower bound on neighborhood size in $G(n,p)$ random graphs (see [7] p. 260).

Lemma 5 (Neighborhood lower bound)

Let ℓ be a fixed constant, If $d \geq n^{1/\ell}(\log(n^2))^{1/\ell}$ and if n is sufficiently large, then

$$\Pr\left[|N^G_{\ell-1}(x)| < \frac{5}{6}(n\log(n^2))^{1-1/\ell}\right] < n^{-4}.$$

Proof of Theorem 8: The $G(n,p)$ model with $p = dn^{-1}$ is a special case of $G(\mathbf{w})$ with $d = \tilde{d} = m$. To prove Theorem 8, notice that in this special case our upper bound from from Lemma 4 becomes

$$\sum_{k\in[1,\ell]} Vol(\Gamma^-_{k-1}(u))Vol(N^L_{\ell-k}(v)) \leq 4(4M)^{\ell-1}d^{\ell+1}.$$

We then obtain

$$\begin{aligned}\Pr\left[|C(u,v)| \geq f \mid B\right] &\leq \rho^f(4(4M)^{\ell-1}d^{\ell+1})^f \\ &= (4(4M)^{\ell-1}d^\ell n)^f,\end{aligned}$$

and the rest of the analysis matches that of the recovery theorem. □

Proof of Theorem 9:

Let $d \geq 6fn^{\frac{1}{\ell}}(\log n)^{\frac{1}{\ell}}$, as in the statement of the theorem. Let u, v be any pair of vertices in H. We will show that R contains f short disjoint paths from u to v, which will imply that every edge in H survives. Partition the vertices $V \setminus \{u,v\}$ into f disjoint sets $V_1 \ldots V_f$, each of size n/f and let H_i be the induced subgraph of H on $V_i \cup \{u,v\}$. We will ignore the fact that we may not be able to partition into sets of size exactly n/f, since it will not be significant. We can view H_i as a $G(n,p)$ random graph with average degree d' satisfying

$$d' = 6n^{\frac{1}{\ell}}(\log n)^{\frac{1}{\ell}} \geq 6n^{\frac{1}{\ell}}(\log n)^{\frac{1}{\ell}} \geq 6|G_i|^{\frac{1}{\ell}}(\log|G_i|)^{\frac{1}{\ell}}.$$

By applying Lemma 5 to any particular G_i,

$$|N^{G_i}_{\ell-1}(x)| \geq \frac{5}{6}(|G_i|\log(|G_i|^2)^{1-1/\ell}$$

with probability at least $1-|G_i|^{-4} = 1-(n/f)^4$. With probability at least $1-f(n/f)^4 \geq 1-n^{-4}$ this holds for all $G_1 \ldots G_f$, and we let A denote this event. If A holds, there is likely to be an edge in G from $N^{G_i}_{\ell-1}(x)$ to v.

$$\begin{aligned}\Pr\left[\text{No edge from } N^{G_i}_{\ell-1}(x) \text{ to } v \mid A\right] &\leq (1-p)^{|N^{G_i}_{\ell-1}(x)|} \\ &\leq exp(-p|N^{G_i}_{\ell-1}(x)|) \\ &\leq exp(-p\frac{5}{6}(|G_i|\log(|G_i|^2)^{1-1/\ell}) \\ &\leq exp(-6fn^{-1}n^{\frac{1}{\ell}}(\log n)^{\frac{1}{\ell}}\frac{5}{6}(|G_i|\log(|G_i|^2)^{1-1/\ell}) \\ &\leq exp(-5\log(n/f)) \\ &\leq O(n^{-5}).\end{aligned}$$

Thus, conditional on A, there is an edge from $N_{\ell-1}^{G_i}(x)$ to v for each $i \in [1, f]$ with probability $1 - fO(n^{-5}) = 1 - O(n^{-5})$. The event A occurs with probability $1 - O(n^{-4})$. Thus, with probability $1 - O(n^4)$ there exist f short disjoint paths from u to v, and hence $f_\ell(u, v) \geq f$. Since there are at most n^2 edges in R, the union bound implies that with probability $1 - O(n^2)$ every edge in R survives. □

8 Probabilistic Analysis of Neighborhoods in Hybrid Graphs

In this section we describe bounds on the growth of neighborhoods in hybrid graphs. Our goal is to prove Lemma 4 by bounding

$$\sum_{k \in [1,\ell]} Vol(\bar{\Gamma}_{k-1}(u)) Vol(N_{\ell-k}^L(v)).$$

The main tool we use is the concentration inequality (8), stated below, which is a generalization of the Chernoff inequalities for the binomial distribution. For a proof, see [12].

Lemma 6 *Let $X_1, \ldots, X_n$ be independent random variables with*

$$Pr(X_i = 1) = p_i, \qquad Pr(X_i = 0) = 1 - p_i$$

For $X = \sum_{i=1}^n a_i X_i$, we let $\mu = E(X) = \sum_{i=1}^n a_i p_i$ and we define $\nu = \sum_{i=1}^n a_i^2 p_i$. Then we have

$$Pr(X < E(X) - \lambda) \leq e^{-\lambda^2/2\nu} \tag{7}$$

$$Pr(X > E(X) + \lambda) \leq e^{-\frac{\lambda^2}{2(\nu + a\lambda/3)}} \tag{8}$$

where $a = \max\{a_1, a_2, \ldots, a_n\}$.

In the remainder of the section we define random variables related to the neighborhoods $\bar{\Gamma}_k(u)$, compute the quantities μ and ν for these random variables, and apply the concentration inequality. We first introduce some notation. Let Q_k again denote the set of vertex pairs

$$Q_k = \bar{\Gamma}_{k-1} \times \left(V \setminus (N_{\ell-k}^L \cup \bar{\Gamma}_0(u) \cup \cdots \cup \bar{\Gamma}_{k-1}(u))\right),$$

and let G_k denote the set of global edges among the pairs Q_k. Let $\bar{\Gamma}_{k,j}(u)$ be the set of vertices $x \in \bar{\Gamma}_k(u)$ such that $x \in N_{k-j}^L(y)$ for some $y \in \bar{\Gamma}_j$, and such that j is the smallest number for which such a y exists. We can think of $\bar{\Gamma}_{k,j}(u)$ as the collection of vertices in $\bar{\Gamma}_k(u)$ which are guaranteed to be in $\bar{\Gamma}_k(u)$ as soon as the edges in G_j are revealed, but not before. We will be considering the volumes of these sets, so we define

$$V_k = Vol(\bar{\Gamma}_k(u)) \quad \text{and} \quad V_{k,j} = Vol(\bar{\Gamma}_{k,j}(u)).$$

The following proposition gives an upper bound on $V_{k,j}$ based on V_{j-1}.

Proposition 3 *Let $\hat{V}_j = \max\{V_j, m\}$, and $\hat{d} = n^\alpha \geq \tilde{d}$. With probability $1 - exp(-\Omega(n^\alpha))$,*

$$V_{k,j} \leq \left(4M^{k-j}\hat{d}\right)\hat{V}_{j-1} \qquad \textit{for all } j \leq k \leq \ell - 1\,.$$

Proof of Proposition 3:

We first make note of the following simple facts:

$$\bar{\Gamma}_k(u) = \bigcup_{j\in[0,k]} \bar{\Gamma}_{j,k}(u) \quad \text{and} \quad V_k \leq \sum_{j\in[0,k]} V_{k,j}.$$

We let $R_{(x,y)}$ be the event that $(x,y) \in R$, and let $\chi(R_{(x,y)})$ be the corresponding indicator random variable. We then rewrite $V_{k,j}$ as

$$\begin{aligned} V_{k,j} &= Vol\left(\bigcup_{\{\, y \mid (x,y)\in G_j\}} \Gamma^L_{k-j}(y)\right) \\ &\leq \sum_{\{y \mid (x,y)\in G_j\}} Vol(\Gamma^L_{k-j}(y)) \\ &\leq \sum_{(x,y)\in Q_j} \chi(R_{(x,y)}) \cdot Vol\left(\Gamma^L_{k-j}(y)\right). \end{aligned}$$

We wish to bound this quantity, so we define the random variable

$$Y_{k,j} = \sum_{(x,y)\in Q_j} \chi(R_{(x,y)}) \cdot Vol\left(\Gamma^L_{k-j}(y)\right).$$

We will use the concentration inequality (8) to bound $Y_{k,j}$, so to that end we compute a, μ, and ν. It is easy to observe that

$$a(Y_{k,j}) = \max_x \left\{Vol\left(\Gamma^L_{k-j}(x)\right)\right\} \leq M^{k-j}m. \tag{9}$$

We then condition on the sets $\bar{\Gamma}_0(u)\ldots\bar{\Gamma}_j(u)$, which determines Q_j, and proceed to compute the expected value of $Y_{k,j}$.

$$\begin{aligned} \mu(Y_{k,j}) &= \sum_{(x,y)\in Q_j} \Pr\left[\,(x,y)\in R\,\right] Vol\left(\Gamma^L_{k-j}(y)\right) \\ &= \sum_{(x,y)\in Q_j} (w_x w_y \rho) Vol\left(\Gamma^L_{k-j}(y)\right). \end{aligned}$$

Although we are conditioning on the sets $\bar{\Gamma}_0(u)\ldots\bar{\Gamma}_j(u)$, the last line follows from Lemma 3 since $Q_j \subseteq S(u,v)$. We now sum over x and y separately, introducing an inequality by summing over all $y \in V$ instead of all $y \in \left(V \setminus (N^L_{\ell-k} \cup \bar{\Gamma}_0(u) \cup \cdots \cup \bar{\Gamma}_{k-1}(u))\right)$.

$$\begin{aligned} \mu(Y_{k,j}) &\leq \sum_{x\in\bar{\Gamma}_{j-1}(u)} \sum_{y\in V} (w_x w_y \rho) Vol\left(\Gamma^L_{k-j}(y)\right) \\ &= \rho\left(\sum_{x\in\bar{\Gamma}_{j-1}(u)} w_x\right) \sum_{y\in V} w_y Vol\left(\Gamma^L_{k-j}(y)\right) \\ &= \rho V_{j-1} \sum_{y\in V} w_y Vol\left(\Gamma^L_{k-j}(y)\right) \\ &= \rho V_{j-1} \sum_{\{(x,y)\,\mid\, y\in V, x\in\Gamma^L_{k-j}(y)\}} w_x w_y. \end{aligned}$$

Finally, we apply Lemma 1 to sum in the last line, noting that each vertex appears in at most $2M^{k-j}$ terms.

$$\begin{aligned} \mu(Y_{k,j}) &\leq \rho V_{j-1} 2M^{k-j} \sum_{y\in V} w_y^2 \\ &= \left(2M^{k-j}\tilde{d}\right) V_{j-1}. \end{aligned}$$

That is the upper bound we will use for μ, and we now compute an upper bound for ν in a similar way.

$$\begin{aligned}\nu(Y_{k,j}) &= \sum_{(x,y)\in Q_j} \Pr\left[(x,y)\in R\right]\cdot Vol\left(\Gamma^L_{k-j}(y)\right)^2 \\ &= \sum_{(x,y)\in Q_j} (w_x w_y \rho)\cdot Vol\left(\Gamma^L_{k-j}(y)\right)^2,\end{aligned}$$

where the last line again follows from Lemma 3 and the fact that $Q_j \subseteq S(u,v)$. Rearranging the sum, we have

$$\begin{aligned}\nu(Y_{k,j}) &\leq \sum_{y\in V}\left(\sum_{x\in\Gamma_{j-1}(u)} w_x w_y \rho\right)\cdot Vol\left(\Gamma^L_{k-j}(y)\right)^2 \\ &= \rho\left(\sum_{x\in\Gamma_{j-1}(u)} w_x\right)\sum_{y\in V} w_y \cdot Vol\left(\Gamma^L_{k-j}(y)\right)^2 \\ &\leq \rho V_{j-1}\sum_{y\in V} w_y\left(\sum_{a\in\Gamma^L_{k-j}(y)}\sum_{b\in\Gamma^L_{k-j}(y)} w_a w_b\right) \\ &= \rho V_{j-1}\left(\sum_{y\in V}\sum_{a\in\Gamma^L_{k-j}(y)}\sum_{b\in\Gamma^L_{k-j}(y)} w_y w_a w_b\right).\end{aligned}$$

Then we apply Lemma 1 to sum in the last line. This time each vertex appears in at most $3(M^{k-j})^2$ terms, and so

$$\begin{aligned}\nu(Y_{k,j}) &\leq \rho V_{j-1}\left(3(M^{k-j})^2\right)\sum_{y\in V} w_y^3 \\ &\leq V_{j-1}\left(3(M^{k-j})^2\right) m \sum_{y\in V} w_y^2\rho \\ &= \left(3(M^{k-j})^2 m\tilde{d}\right) V_{j-1}.\end{aligned}$$

We are ready to combine these results and apply the concentration inequality. Recall from the statement of Proposition 3 that $\hat{V}_j = \max\{V_j, m\}$. We define $\hat{\mu}_{k,j} = \left(2M^{k-j}\hat{d}\right)\hat{V}_{j-1}$, and note that

$$\hat{\mu}_{k,j} = \left(2M^{k-j}\hat{d}\right)\hat{V}_{j-1} \geq \left(2M^{k-j}\tilde{d}\right)V_{j-1} \geq \mu(Y_{k,j}).$$

Therefore,

$$\Pr\left[Y_{k,j} > 2\hat{\mu}_{k,j}\right] \leq \Pr\left[Y_{k,j} > \mu(Y_{k,j}) + \lambda\right]$$

for some λ satisfying $\hat{\mu}_{k,j} \leq \lambda \leq 2\hat{\mu}_{k,j}$.

Applying the concentration inequality,

$$\begin{aligned}
\Pr\left[Y_{k,j} > 2\hat{\mu}_{k,j}\right] &\leq exp\left(-\frac{\hat{\mu}_{k,j}^2}{2(\nu + \frac{a(2\hat{\mu}_{k,j})}{3})}\right) \\
&\leq exp\left(-\frac{\left(2M^{k-j}\hat{d}\right)^2 (\hat{V}_{j-1})^2}{2\left(\left(3(M^{k-j})^2 m\tilde{d}\right) V_{j-1} + \frac{2\left(2M^{k-j}\hat{d}\right)\hat{V}_{j-1}(M^{k-j}m)}{3}\right)}\right) \\
&\leq exp\left(-\frac{4\hat{V}_{j-1}}{2\left(3m + \frac{4m}{3}\right)}\hat{d}\right) \\
&\leq exp\left(-\frac{6}{13}\frac{\hat{V}_{j-1}}{m}\hat{d}\right) \\
&\leq exp\left(-\Omega(\hat{d})\right) \\
&= exp\left(-\Omega(n^\alpha)\right).
\end{aligned}$$

Thus we have that for any fixed k, j, the following holds with probability $1 - e^{(-\Omega(n^\alpha))}$:

$$Y_{k,j} \leq 2\hat{\mu}_{k,j} = \left(4M^{k-j}\hat{d}\right)\hat{V}_{j-1}.$$

The union bound implies that this holds for all $j \leq k \leq \ell - 1$ with probability $1 - \ell^2 e^{-\Omega(n^\alpha)} = 1 - e^{-\Omega(n^\alpha)}$. This completes the proof of Proposition 3. □

Proposition 4 *With probability* $1 - exp(-\Omega(n^\alpha))$,

$$V_k \leq (4M\hat{d})^k m \qquad \textit{for all } k \in [0, \ell - 1].$$

Proof of Proposition 4: We prove by induction that

$$V_k \leq (4M\hat{d})^k m, \tag{10}$$

given that

$$V_{k,j} \leq \left(4M^{k-j}\hat{d}\right)\hat{V}_{j-1} \qquad \text{for all } j \leq k \leq \ell - 1.$$

The result of Proposition 4 will follow immediately, since that event occurs with probability $1 - exp(-\Omega(n^\alpha))$ by Proposition 3.

Equation (10) holds for $k = 0$ since we have $V_0 = Vol(\{u\}) \leq m$. Assume now that (10) holds for $[0, k]$ and consider V_{k+1}.

$$
\begin{aligned}
V_{k+1} &\leq \sum_{j\in[0,k+1]} Y_{k+1,j} \\
&\leq Y_{k+1,0} + \sum_{j\in[1,k+1]} \left(4M^{k+1-j}\hat{d}\right)\hat{V}_{j-1} \\
&\leq \Gamma^L_{k+1}(u) + \sum_{j\in[0,k+1]} \left(4M^{k+1-j}\hat{d}\right)(4M\hat{d})^{j-1}m \\
&\leq M^{k+1} + M^k m \sum_{j\in[0,k+1]} \left(4\hat{d}\right)^j \\
&\leq M^{k+1}m(4\hat{d})^{(k+1)} \\
&= (4M\hat{d})^{k+1}m.
\end{aligned}
$$

In the second-to-last line we have assumed that $M \geq 2$. This completes the proof of Proposition 4. □

Proof of Lemma 4:

With probability $1 - e^{-\Omega(n^\alpha)}$,

$$
\begin{aligned}
\sum_{k\in[1,\ell]} Vol(\bar{\Gamma}_{k-1}(u))Vol(N^L_{\ell-k}(v)) &\leq \sum_{k\in[1,\ell]} \left((4M\hat{d})^{k-1}m\right)(2mM^{\ell-k}) \\
&= 2m^2M^{\ell-1}\sum_{k\in[1,\ell]}(4\hat{d})^{k-1} \\
&\leq 4m^2M^{\ell-1}(4\hat{d})^{\ell-1} \\
&= 4m^2(4M\hat{d})^{\ell-1}
\end{aligned}
$$

□

9 Communities and Examples

The local graph L found by the `Extract`(f, l) algorithm is not necessarily connected. Each connected component of L can be viewed as a local community. By fixing l and increasing f we obtain a hierarchy of successively smaller communities.

Flake et al. [17] defined a hierarchy of communities using minimum cut trees. Their communities have provably good expansion, and few edges between communities. The communities found by `Extract` are highly locally connected, are robust to the addition of random edges, and are are monotone in the sense that adding edges can only increase the size of a community. These communities often have rich structures other than cliques or complete bipartite subgraphs.

We applied the `Recursive Extract` algorithm to a routing graph G collected by "champagne.sdsc.org". The maximum 3-connected subgraph of G consists of 7 copies of K_4 and a large connected component L with 2364 vertices and 5947 edges. Applying `Recursive Extract` with parameters $(f = 3, \ell = 3)$ breaks L into 79 non-singleton communities. The largest community has 881 vertices. The second largest community (of size 59) is illustrated in Figure 5, and two communities of size 25 and 35 are illustrated in Figures 6 and 7.

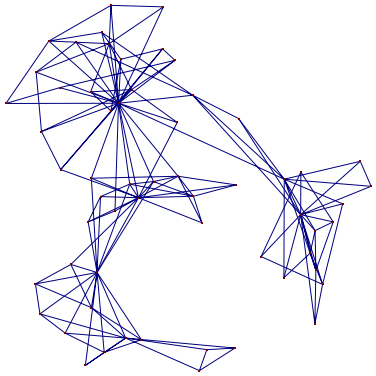

Figure 5: *A (3,3)-connected community of size 59 in a routing graph.*

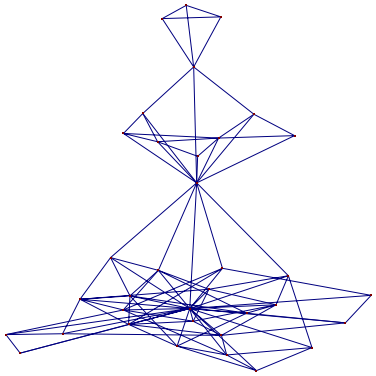

Figure 6: *A (3,3)-connected community of size 35 in the routing graph.*

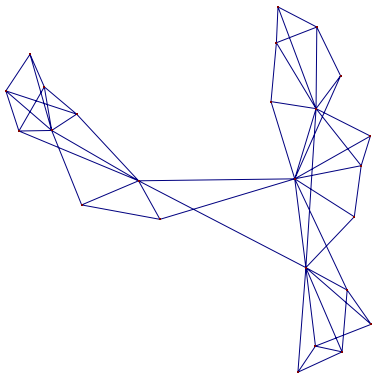

Figure 7: *A (3,3)-connected community of size 25 in the routing graph.*

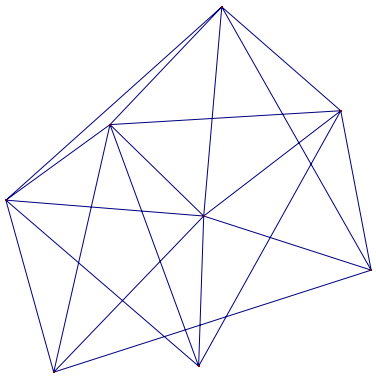

Figure 8: *A* $(4,3)$*-connected sub-community of the community in Figure 6.*

References

[1] L. A. Adamic and B. A. Huberman, Growth dynamics of the World Wide Web, *Nature*, **401**, September 9, 1999, pp. 131.

[2] W. Aiello, F. Chung and L. Lu, A random graph model for massive graphs, *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, (2000) 171-180.

[3] R. B. R. Azevedo and A. M. Leroi, A power law for cells, *Proc. Natl. Acad. Sci. USA*, vol. **98**, no. 10, (2001), 5699-5704.

[4] Georg Baier, Flows with Path Restrictions. PhD thesis, Technische Universität Berlin, 2003.

[5] Albert-László Barabási and Réka Albert, Emergence of scaling in random networks, Science **286** (1999) 509-512.

[6] A. Barabási, R. Albert, and H. Jeong, Scale-free characteristics of random networks: the topology of the world wide web, *Physica* A 272 (1999), 173-187.

[7] B. Bollobás, *Random Graphs*, Academic, New York, 1985.

[8] S. Boyles, G. Exoo, On line disjoint paths of bounded length. *Discrete Math.* **44** (1983)

[9] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tompkins, and J. Wiener, "Graph Structure in the Web," *proceedings of the WWW9 Conference*, May, 2000, Amsterdam. Paper version appeared in *Computer Networks* **33**, (1-6), (2000), 309-321.

[10] K. Calvert, M. Doar, and E. Zegura, Modeling Internet topology. *IEEE Communications Magazine*, **35(6)** (1997) 160-163.

[11] F. Chung and L. Lu, Average distances in random graphs with given expected degree sequences, *Proceedings of National Academy of Science*, **99** (2002).

[12] Fan Chung and Linyuan Lu, Connected components in a random graph with given degree sequences, *Annals of Combinatorics*, **6** (2002), 125-145.

[13] F. Chung and L. Lu, The small world phenomenon in hybrid power law graphs, *Complex Networks*, (Eds. Eli Ben-Naim, Hans Frauenfelder and Zoltan Toroczk ai), *Lecture Notes in Physics*, Vol. 650, Springer-Verlag, (2004), 91-106.

[14] C. Cooper and A. Frieze, On a general model of web graphs, *Random Structures and Algorithms* Vol. **22**, (2003), 311-335.

[15] M. Faloutsos, P. Faloutsos, and C. Faloutsos, On power-law relationships of the Internet topology, *Proceedings of the ACM SIGCOM Conference*, Cambridge, MA, 1999.

[16] P. Erdős and T. Gallai, Gráfok előírt fokú pontokkal (Graphs with points of prescribed degrees, in Hungarian), *Mat. Lapok* **11** (1961), 264-274.

[17] Gary William Flake, Robert E. Tarjan and Kostas Tsioutsiouliklis Graph Clustering and Minimum Cut Trees Internet Math. 1, no. 4 (2004), 385-408

[18] N. Garg, J. Könemann, Faster and simpler algorithms for multicommodity flow and other fractional packing problems. *Technical Report, Max-Planck-Institut fur Informatik, Saarbrucken, Germany* (1997).

[19] A. Itai, Y. Perl, and Y. Shiloach, The complexity of finding maximum disjoint paths with length constraints, *Networks* **12** (1982)

[20] S. Jain and S. Krishna, A model for the emergence of cooperation, interdependence, and structure in evolving networks, *Proc. Natl. Acad. Sci. USA*, vol. **98**, no. 2, (2001), 543-547.

[21] J. Kleinberg, The small-world phenomenon: An algorithmic perspective, *Proc. 32nd ACM Symposium on Theory of Computing*, 2000.

[22] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins and E. Upfal, Stochastic models for the web graph, Proceedings of the 41st IEEE Symposium on Foundations of Computer Science (2000).

[23] S. R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, Extracting large-scale knowledge bases from the web, *Proceedings of the 25th VLDB Conference*, Edinburgh, Scotland, 1999.

[24] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins and E. Upfal, The web as a graph, Proceedings of the 19th ACM Symposium on Principles of Database Systems (2000).

[25] L. Lovász, V. Neumann-Lara, and M. Plummer. Mengerian the- orems for paths of bounded length. Periodica Mathematica Hun- garica, 9:269276, 1978.

[26] M. Mitzenmacher, A Brief History of Generative Models for Power Law and Lognormal Distributions, *Internet Math.* 1 (2003), no. 2.

[27] S. Plotkin, D. B. Shmoys, and E Tardos, Fast approximation algorithms for fractional packing and covering problems, *FOCS* 1991, pp. 495–504.

[28] M. E. J., Newman, The structure of scientific collaboration networks, *Proc. Natl. Acad. Sci. USA*, vol. **98**, no. 2, (2001), 404-409.

[29] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small small-world networks, *Nature* **393**, 440-442.

[30] Neal E. Young, Randomized rounding without solving the linear program, Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms (1995).