Discovery through Gossip

Bernhard Haeupler [*] Gopal Pandurangan [†] David Peleg [‡] Rajmohan Rajaraman [§]
Zhifeng Sun [§]

Abstract

We study stochastic processes in dynamic networks that are motivated by information discovery in large-scale distributed networks such as peer-to-peer and social networks. A well-studied problem in peer-to-peer networks is *resource discovery*, where the goal for nodes (hosts with IP addresses) is to discover the IP addresses of all other hosts. In social networks, nodes (people) discover new nodes through exchanging contacts with their neighbors (friends). In both cases the discovery of new nodes changes the underlying network — new edges are added to the network — and the process continues in the changed network.

This paper studies and analyzes two natural gossip-based discovery processes. In the *push discovery* or *triangulation* process, each node repeatedly chooses two random neighbors and connects them (i.e., "pushes" their mutual information to each other). In the *pull discovery* process or the *two-hop walk*, each node repeatedly requests or "pulls" a random contact from a random neighbor and connects itself to this two-hop neighbor. Both processes are lightweight in the sense that the amortized work done per node is constant per round, local, and naturally robust due to the inherent randomized nature of gossip.

Our main result is an almost-tight analysis of the time taken for these two randomized processes to converge. We show that in any undirected n-node graph both processes take $O(n \log^2 n)$ rounds to connect every node to all other nodes with high probability, whereas $\Omega(n \log n)$ is a lower bound. We also study the two-hop walk in directed graphs, and show that it takes $O(n^2 \log n)$ time with high probability, and that the worst-case bound is tight for arbitrary directed graphs, whereas $\Omega(n^2)$ is a lower bound for strongly connected directed graphs. A key technical challenge that we overcome in our work is the analysis of a randomized process that itself results in a constantly changing network leading to complicated dependencies in every round.

Keywords: Random process, Resource discovery, Social network, Gossip-based algorithm, Distributed algorithm, Probabilistic analysis

[*]Computer Science and Artificial Intelligence Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. E-mail: haeupler@mit.edu

[†]Division of Mathematical Sciences, Nanyang Technological University, Singapore 637371 and Department of Computer Science, Brown University, Providence, RI 02912, USA. E-mail: gopalpandurangan@gmail.com. Supported in part by the following grants: Nanyang Technological University grant M58110000, Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 2 grant MOE2010-T2-2-082, US NSF grant CCF-1023166, and a grant from the US-Israel Binational Science Foundation (BSF).

[‡]Department of Computer Science and Applied Mathematics, The Weizmann Institute of Science, Rehovot, 76100 Israel. E-mail: david.peleg@weizmann.ac.il. Supported by a grant from the United States-Israel Binational Science Foundation (BSF).

[§]College of Computing and Information Science, Northeastern University, Boston MA 02115, USA. E-mail: {rraj,austin}@ccs.neu.edu.

1 Introduction

Many large-scale, real-world networks such as peer-to-peer networks, the Web, and social networks are highly dynamic with continuously changing topologies. The evolution of the network as a whole is typically determined by the decentralized behavior of nodes, i.e., the local topological changes made by the individual nodes (e.g., adding edges between neighbors). Understanding the dynamics of such local processes is critical for both analyzing the underlying stochastic phenomena, e.g., in evolution of social networks, the Web and other real-world networks [4, 20, 21], and designing practical algorithms for associated algorithmic problems, e.g., in resource discovery in distributed networks [11, 17] or in the analysis of algorithms for the Web [6, 8]. In this paper, we study the dynamics of network evolution that result from *local* gossip-style processes. Gossip-based processes have recently received significant attention because of their simplicity of implementation, scalability to large network size, and robustness to frequent network topology changes; see, e.g., [9, 14, 15, 7, 13, 12, 19, 5] and the references therein. In a local gossip-based algorithm (e.g., [7]), each node exchanges information with a small number of randomly chosen neighbors in each round.[1] The randomness inherent in the gossip-based protocols naturally provides robustness, simplicity, and scalability.

We present two illustrative applications for our study. First, consider a P2P network, where nodes (computers or end-hosts with IDs/IP addresses) can communicate only with nodes whose IP address are known to them. A basic building block of such a dynamic distributed network is to efficiently discover the IP addresses of all nodes that currently exist in the network. This task, called *resource discovery* [11], is a vital mechanism in a dynamic distributed network with many applications [11, 1]: when many nodes in the system want to interact and cooperate they need a mechanism to discover the existence of one another. Resource discovery is typically done using a local mechanism [11]; in each *round* nodes discover other nodes and this changes the resulting network — new edges are added between the nodes that discovered each other. As the process proceeds, the graph becomes denser and denser and will finally result in a complete graph. Such a process was first studied in [11] which showed that a simple randomized process is enough to guarantee almost-optimal time bounds for the time taken for the entire graph to become complete (i.e., for all nodes to discover all other nodes). Their randomized *Name Dropper* algorithm operates as follows: in each round, each node chooses a random neighbor and sends *all* the IP addresses it knows. Note that while this process is also gossip based the information sent by a node to its neighbor can be extremely large (i.e., of size $\Omega(n)$).

Second, in social networks, nodes (people) discover new nodes through exchanging contacts with their neighbors (friends). Discovery of new nodes changes the underlying network — new edges are added to the network — and the process continues in the changed network. For example, consider the *LinkedIn* network[2], a large social network of professionals on the Web. The nodes of the network represent people and edges are added between people who directly know each other — between direct contacts. Edges are generally undirected, but LinkedIn also allows directed edges, where only one node is in the contact list of another node. LinkedIn allows two mechanisms to discover new contacts. The first can be thought of as a *triangulation* process (see Figure 1(a)): A person can introduce two of his friends that could benefit from knowing each other — he can mutually introduce them by giving their contacts. The second can be thought of as a *two-hop* process (see Figure 1(b)): If *you* want to acquire a new contact then you can use a shared (mutual) neighbor to introduce yourself to this contact; i.e., the new contact has to be a two-hop neighbor of yours. Both the processes can be modeled via gossip in a natural way and the resulting evolution of the network can be studied. This yields insight on the evolution of the social network over time.

[1] Gossip, in some contexts (see e.g., [12, 13]), has been used to denote communication with a random node in the network, as opposed to only a directly connected neighbor. The former model essentially assumes that the underlying graph is complete, whereas the latter (as assumed here) is more general and applies even to arbitrary graphs. The local gossip process is typically more difficult to analyze due to the dependences that arise as the network evolves.

[2] `http://www.linkedin.com`.

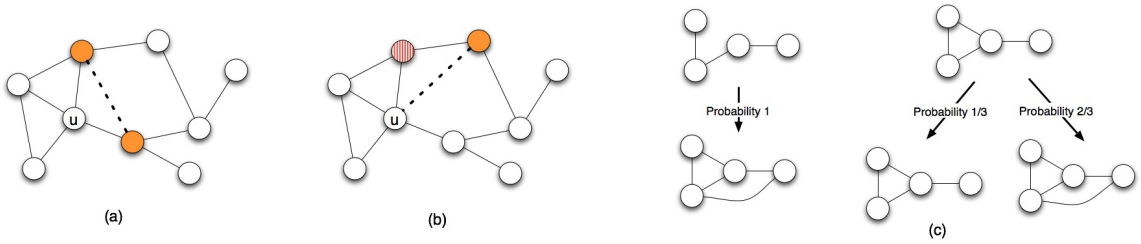

Figure 1: (a) Push discovery or triangulation process. (b) Pull discovery or two-hop walk process. (c) Non-monotonicity of the triangulation process – the expected convergence time for the 4-edge graph exceeds that for the 3-edge subgraph.

Gossip-based discovery. Motivated by the above applications, we analyze two natural gossip-based discovery processes. We assume that we start with an arbitrary undirected connected graph and the process proceeds in synchronous rounds. Communication among nodes occurs only through edges in the network. We further assume that the size of each message sent by a node in a round is at most $O(\log n)$ bits, i.e., the size of an ID.

1. Push discovery (triangulation): In each round, each node chooses two random neighbors and connects them by "pushing" their mutual information to each other. In other words, each node adds an undirected edge between two of its random neighbors; if the two neighbors are already connected, then this does not create any new edge. Note that this process, which is illustrated in Figure 1(a), is completely local. To execute the process, a node only needs to know its neighbors; in particular, no two-hop information is needed.

2. Pull discovery (two-hop walk): In each round, each node connects itself to a random neighbor of one of its randomly chosen neighbors, by "pulling" a random neighboring ID from a random neighbor. Alternatively, one can think of each node doing a two-hop random walk and connecting to its destination. This process, illustrated in Figure 1(b), can also be executed locally: a node simply asks one of its neighbors v for an ID of one of v's neighbors and then adds an undirected edge to the received contact.

Both the above processes are local in the sense that each node only communicates with its neighbors in any round, and lightweight in the sense that the amortized work done per node is only a constant per round. Both processes are also easy to implement and generally oblivious to the current topology structure, changes or failures. It is interesting also to consider variants of the above processes in directed graphs. In particular, we study the two-hop walk process which naturally generalizes in directed graphs: each node does a two-hop directed random walk and adds a *directed* edge to its destination. We are mainly interested in the time taken by the process to converge to the transitive closure of the initial graph, i.e., till no more new edges can be added.

Our results. We present almost-tight bounds on the number of rounds it takes for the push and pull discovery processes to converge.

- **Undirected graphs:** In Sections 3 and 4, we show that for *any* undirected n-node graph, both the push and the pull discovery processes converge in $O(n \log^2 n)$ rounds with high probability. We also show that $\Omega(n \log n)$ is a lower bound on the number of rounds needed for almost any n-node graph. Hence our analysis is tight to within a logarithmic factor.

- **Directed graphs:** In Section 5, we show that the pull process takes $O(n^2 \log n)$ time for any n-node directed graph, with high probability. We show a matching lower bound for weakly connected graphs, and an $\Omega(n^2)$ lower bound for strongly connected directed graphs. Our analysis indicates that the directionality of edges can greatly impede the resource discovery process.

Applications. The gossip-based discovery processes we study are directly motivated by the two scenarios outlined above, namely algorithms for resource discovery in distributed networks and analyzing how discovery process affects the evolution of social networks. Since our processes are simple, lightweight, and easy to implement, they can be used for resource discovery in distributed networks. The original resource discovery algorithm of [11] was helpful in developing systems like Akamai. Unlike prior algorithms for the discovery problem [11, 17, 16, 1], the amortized work done per node in our processes is only constant per round and hence this can be efficiently implemented in bandwidth and resource-constrained networks (e.g., peer-to-peer or sensor networks). In contrast, the *Name Dropper* algorithm of [11], can transfer up to $\Theta(n)$ information per edge per round and hence may not be scalable for large-scale networks. We note that, however, because there is essentially no restriction on the bandwidth, the number of rounds taken by the *Name Dropper* algorithm is $O(\log^2 n)$. (We note that in our model, $\Omega(n)$ is a trivial lower bound). Our analyses can also give insight into the growth of real-social networks such as LinkedIn, Twitter, or Facebook, that grow in a decentralized way by the local actions of the individual nodes. For example, it can help in predicting the sizes of the immediate neighbors as well as the sizes of the second and third-degree neighbors (e.g., these are listed for every node in LinkedIn). An estimate of these can help in designing efficient algorithms and data structures to search and navigate the social network.

Technical contributions. Our main technical contribution is a probabilistic analysis of localized gossip-based discovery in arbitrary networks. While our processes can be viewed as graph-based coupon collection processes, one significant distinction with past work in this area [2, 3, 10] is that the graphs in our processes are constantly changing. The dynamics and locality inherent in our process introduces nontrivial dependencies, which makes it difficult to characterize the network as it evolves. A further challenge is posed by the fact that the expected convergence time for the two processes is *not monotonic*; that is, the processes may *take longer* to converge starting from a graph G than starting from a subgraph H of G. Figure 1(c) presents a small example illustrating this phenomenon. This seemingly counterintuitive phenomenon is, however, not surprising considering the fact that the cover time of random walks also share a similar property. One consequence of these hurdles is that analyzing the convergence time for even highly specialized or regular graphs is challenging since the probability distributions of the intermediate graphs are hard to specify. Our lower bound analysis for a specific strongly connected directed graph in Theorem 15 illustrates some of the challenges. In our main upper bound results, we overcome these technical difficulties by presenting a uniform analysis for all graphs, in which we study different local neighborhood structures and show how each lead to rapid growth in the minimum degree of the graph.

2 Preliminaries

In this section, we define the notations used in our proofs, and prove some common lemmas for Section 3 and Section 4. Let G denote a connected graph, $d(u)$ denote the degree of node u, and $N^i(u)$ denote the set of nodes that are at distance i from u. Let δ denote the minimum degree of G. We note that G, $d(u)$, and $N^i(u)$ all change with time, and are, in fact, random variables. For any nonnegative integer t, we use subscript t to denote the random variable at the start of time t; for example G_t refers to the graph at the start of step t. For convenience, we list the notations in Table 1.

Table 1: Notation table

Notation	description
δ_t	minimum degree of graph G_t
$N_t^i(u)$	set of nodes that are at distance i from u in G_t
$\left\|N_t^i(u)\right\|$	number of nodes in $N_t^i(u)$
$d_t(u)$	degree of node u in G_t
$d_t\left(u, N_t^i(v)\right)$	number of edges from u to nodes in $N_t^i(v)$, i.e., degree induced on $N_t^i(v)$

We present two lemmas that are used in the proofs in Section 3 and Section 4. Lemma 1 gives a lower bound on the number of neighbors within distance 4 for any node u in G_t while Lemma 2 is a standard analysis of a sequence of Bernoulli experiments.

Lemma 1. $\left|\cup_{i=1}^4 N_t^i(u)\right| \geq \min\{2\delta_t, n-1\}$ *for all* u *in* G_t.

Proof. If $N_t^3(u)$ is not an empty set, consider node $v \in N_t^3(u)$. Since $d_t(v) \geq \delta_t$, we have $\left|\cup_{i=2}^4 N_t^i(u)\right| \geq \delta_t$. $\left|N_t^1(u)\right| \geq \delta_t$ because $d_t(u) \geq \delta_t$. We also know $N_t^1(u)$ and $\cup_{i=2}^4 N_t^i(u)$ are disjoint. Thus, $\left|\cup_{i=1}^4 N_t^i(u)\right| \geq 2\delta_t$. If $N_t^3(u)$ is an empty set, then $N_t^1(u) \cup N_t^2(u) = n-1$ because G_t is connected. Thus $\left|\cup_{i=1}^4 N_t^i(u)\right| = n-1$. Combine the above 2 cases, we complete the proof of this lemma. □

Lemma 2. *Consider* k *Bernoulli experiments, in which the success probability of the* i*th experiment is at least* i/m *where* $m \geq k$*. If* X_i *denotes the number of trials needed for experiment* i *to output a success and* $X = \sum_{i=1}^k X_i$*, then*

$$\Pr[X > (c+1)n\ln n] < \frac{1}{n^c}$$

Proof. Since X only increases with k, with out loss of generality assume that $k = m$. Now we can view this as *coupon collector problem* [18] where X_{m+1-i} is the number of steps to collect the ith coupon. Consider the probability of not obtaining the ith coupon after $(c+1)n\ln n$ steps. This probability is

$$\left(1 - \frac{1}{n}\right)^{(c+1)n\ln n} < e^{-(c+1)\ln n} = \frac{1}{n^{c+1}}$$

By union bound, the probability that some coupon has not been collected after $(c+1)n\ln n$ steps is less than $1/n^c$. And this completes the proof of this lemma. □

3 The triangulation: Discovery through push

In this section, we analyze the triangulation process on undirected connected graphs, which is described by the following simple iteration: In each round, for each node u, we add edge (v, w) where v and w are drawn uniformly at random from $N_t^1(u)$. The triangulation process yields the following push-based resource discovery protocol. In each round, each node u introduces two random neighbors v and w to one another. The main result of this section is that the triangulation process transforms an arbitrary connected n-node graph to a complete graph in $O(n\log^2 n)$ rounds with high probability. We also establish an $\Omega(n\log n)$ lower bound on the triangulation process for almost all n-node graphs.

3.1 Upper bound

We obtain the $O(n \log^2 n)$ upper bound by proving that the minimum degree of the graph increases by a constant factor (or equals $n-1$) in $O(n \log n)$ steps. Towards this objective, we study how the neighbors of a given node connect to the two-hop neighbors of the node. We say that a node v is ***weakly tied*** to a set of nodes S if v has less than $\delta_0/2$ edges to S (i.e., $d_t(v, S) < \delta_0/2$), and ***strongly tied*** to S if v has at least $\delta_0/2$ edges to S (i.e., $d_t(v, S) \geq \delta_0/2$). Recall that δ_0 is the minimum degree at start of round 0. Then, we have the following two lemmas.

Lemma 3. *If $\delta_0 \leq d_t(u) < (1+1/4)\delta_0$ and $w \in N_0^1(u)$ is strongly tied to $N_t^2(u)$, then the probability that u connects to a node in $N_t^2(u)$ through w in round t is at least $2/(7n)$.*

Proof. Since w is strongly tied to $N_t^2(u)$, $d_t\left(w, N_t^2(u)\right) \geq \delta_0/2$. Therefore, the probability that u connects to a node in $N_t^2(u)$ through w in round t is

$$= \frac{d_t\left(w, N_t^2(u)\right)}{d_t(w)} \cdot \frac{1}{d_t(w)} \geq \frac{d_t\left(w, N_t^2(u)\right)}{d_t(w)} \cdot \frac{1}{n} \geq \frac{d_t\left(w, N_t^2(u)\right)}{|N_t^1(u)| + d_t\left(w, N_t^2(u)\right)} \cdot \frac{1}{n}$$

$$\geq \frac{d_t\left(w, N_t^2(u)\right)}{(1+1/4)\delta_0 + d_t\left(w, N_t^2(u)\right)} \cdot \frac{1}{n} \geq \frac{\delta_0/2}{(1+1/4)\delta_0 + \delta_0/2} \cdot \frac{1}{n} = \frac{2}{7n}.$$

□

Lemma 4. *If $\delta_0 \leq d_t(u) < (1+1/4)\delta_0$, $w \in N_0^1(u)$ is weakly tied to $N_t^2(u)$, and $v \in N_0^2(u) \cap N_0^1(w)$, then the probability that u connects to v through w in round t is at least $1/(4\delta_0^2)$.*

Proof. Since w is weakly tied to $N_t^2(u)$, we know that $d_t(w)$ equals $|N_t^1(u)| + d_t\left(w, N_t^2(u)\right)$, which is at most $(1+1/4)\delta_0 + \delta_0/2$. Therefore, the probability that u connects to v through w in round t is

$$= \frac{1}{d_t(w)^2} \geq \frac{1}{((1+1/4)\delta_0 + \delta_0/2)^2} \geq \frac{1}{(7\delta_0/4)^2} \geq \frac{1}{4\delta_0^2}.$$

□

For analyzing the growth in the degree of a node u, we consider two overlapping cases. The first case is when more than $\delta_0/4$ nodes of $N_t^1(u)$ are strongly tied to $N_t^2(u)$, and the second is when less than $\delta_0/3$ nodes of $N_t^1(u)$ are strongly tied to $N_t^2(u)$. The analysis for the first case is relatively straightforward: when several neighbors of a node u are strongly tied to u's two-hop neighbors, then their triangulation steps connect u to a large fraction of these two-hop neighbors.

Lemma 5 (When several neighbors are strongly tied to two-hop neighbors). *There exists $T = O(n \log n)$ such that if more than $\delta_0/4$ nodes in $N_t^1(u)$ are strongly tied to $N_t^2(u)$ for all $t < T$, then $d_T(u) \geq (1+1/4)\delta_0$ with probability at least $1 - 1/n^2$.*

Proof. If at any round $t < T$, $d_t(u) \geq (1+1/4)\,\delta_0$, then the claim of the lemma holds. In the remainder of this proof, we assume $d_t(u) < (1+1/4)\,\delta_0$ for all $t < T$. Let $w \in N_t^1(u)$ be a node that is strongly tied to $N_t^2(u)$. By Lemma 3 we know that

$$\Pr\left[u \text{ connects to a node in } N_t^2(u) \text{ through } w \text{ in round } t\right] \geq \frac{2}{7n} > \frac{1}{6n}$$

We have more than $\delta_0/4$ such w's in $N_t^1(u)$, each of which independently executes a triangulation step in any given round. Consider a run of $T_1 = 72n \ln n/\delta_0$ rounds. This implies at least $18n \ln n$ attempts to add an edge between u and a node in $N_t^2(u)$. Thus,

$$\Pr\left[u \text{ connects to a node in } N_t^2(u) \text{ after } T_1 \text{ rounds}\right] \geq 1 - \left(1 - \frac{1}{6n}\right)^{18n \ln n} \geq 1 - e^{-3 \ln n} = 1 - \frac{1}{n^3}.$$

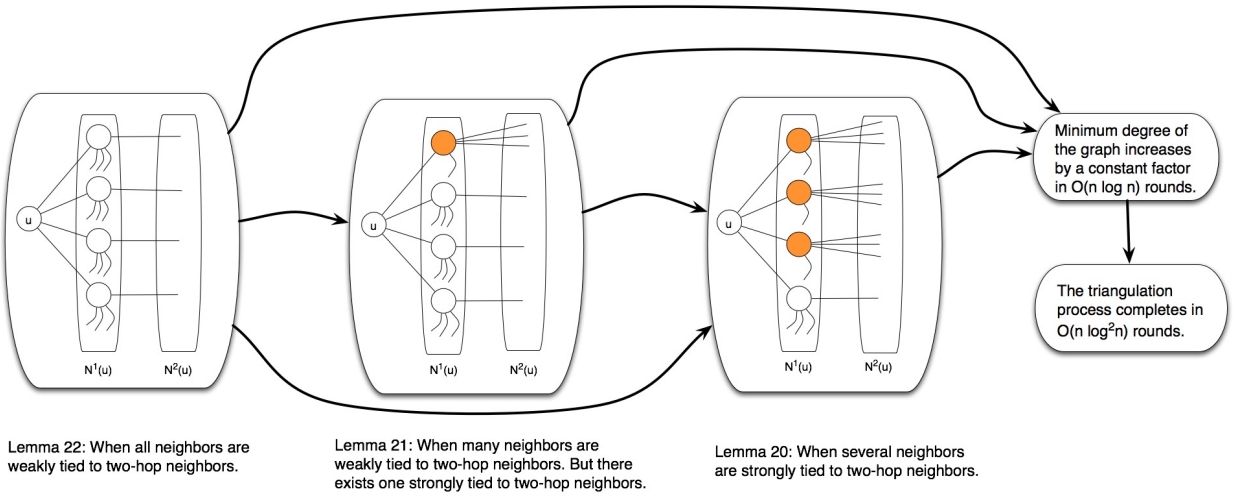

Figure 2: This figure illustrates the different cases and relations between lemmas used in the proof of Theorem 8. The shaded nodes in $N_t^1(u)$ are strongly tied to $N_t^2(u)$. Others are weakly tied to $N_t^2(u)$.

Therefore, in $T = T_1\delta_0/4 = O(n \log n)$ rounds, u will connect to at least $\delta_0/4$ new nodes with probability at least $1 - 1/n^2$, i.e., $d_T(u) \geq (1 + 1/4)\,\delta_0$. □

We next consider the second case where less than $\delta_0/3$ neighbors of a given node u are strongly tied to the two-hop neighborhood of u. This case is more challenging since the neighbors of u that are weakly tied may not contribute many new edges to u. We break the analysis of this part into two subcases based on whether there is at least one neighbor of u that is strongly tied to $N_0^2(u)$.

Lemma 6 (When few neighbors are strongly tied to two-hop neighbors). *There exists $T = O(n \log n)$ such that if less than $\delta_0/3$ nodes in $N_t^1(u)$ are strongly tied to $N_t^2(u)$ for all $t < T$, and there exists a node $v_0 \in N_0^1(u)$ that is strongly tied to $N_0^2(u)$, then $d_T(u) \geq (1 + 1/8)\,\delta_0$ with probability at least $1 - 1/n^2$.*

Proof. If at any point $t < T$, $d_T(u) \geq (1 + 1/8)\,\delta_0$, then the claim of the lemma holds. In the remainder of this proof, we assume $d_T(u) < (1 + 1/8)\,\delta_0$ for all $t < T$. Let S_t^0 denote the set of v_0's neighbors in $N_t^2(u)$ which are strongly tied to $N_t^1(u)$ at time t, W_t^0 denote the set of v_0's neighbors in $N_t^2(u)$ which are weakly tied to $N_t^1(u)$ at time t.

Consider any node v in S_t^0. Less than $\delta_0/3$ nodes in $N_t^1(u)$ are strongly tied to $N_t^2(u)$, thus more than $\delta_0/2 - \delta_0/3 = \delta_0/6$ neighbors of v in $N_t^1(u)$ are weakly tied to $N_t^2(u)$. Let w be one such weakly tied node. By Lemma 4, the probability that u connects to v through w in round t is at least $1/(4\delta_0^2)$. We have at least $\delta_0/6$ such w's, each of which executes a triangulation step each round. Consider $T = 72\delta_0 \ln n$ rounds of the process. Then the probability that u connects to v in T rounds is at least

$$1 - \left(1 - \frac{1}{4\delta_0^2}\right)^{12\delta_0^2 \ln n} \geq 1 - e^{-3\ln n} = 1 - \frac{1}{n^3}.$$

Thus, if $|S_t^0| \geq \delta_0/8$, in an additional $O(n \log n)$ time, $d_T(u) \geq (1 + 1/8)\delta_0$ with probability at least $1 - 1/n^2$.

Therefore, in the remainder of the proof we consider the case where $|S_t^0| < \delta_0/8$. Define $R_t^0 = R_{t-1}^0 \cup W_t^0$, $R_0^0 = W_0^0$. If at least $\delta_0/8$ nodes in R_t^0 are connected to u at any time, then the claim of the lemma

holds. Thus, in the following we consider the case where $|R_t^0 \cap N_t^1(u)| < \delta_0/8$. From the definition of R_t^0, we can derive

$$|R_t^0| \geq |W_t^0| = d_t\left(v_0, N_t^2(u)\right) - |S_t^0| \geq d_t\left(v_0, N_t^2(u)\right) - \delta_0/8$$

At time 0, v_0 is strongly tied to $N_0^2(u)$, i.e., $d_0\left(v_0, N_0^2(u)\right) \geq \delta_0/2$. Since $\delta_0 \leq d_t(u) < (1+1/8)\delta_0$, we have

$$d_t\left(v_0, N_t^2(u)\right) \geq d_t\left(v_0, N_0^2(u)\right) - \delta_0/8 \geq 3\delta_0/8$$

Let e_1 denote the event $\left\{u \text{ connects to a node in } R_t^0 \setminus N_t^1(u) \text{ through } v_0 \text{ in round } t\right\}$.

$$\begin{aligned}
\Pr[e_1] &= \frac{|R_t^0 \setminus N_t^1(u)|}{d_t(v_0)} \cdot \frac{1}{d_t(v_0)} = \frac{|R_t^0| - |R_t^0 \cap N_t^1(u)|}{d_t(v_0)} \cdot \frac{1}{d_t(v_0)} \\
&\geq \frac{|R_t^0| - |R_t^0 \cap N_t^1(u)|}{d_t(v_0)} \cdot \frac{1}{n} = \frac{|R_t^0| - |R_t^0 \cap N_t^1(u)|}{\left|N_t^1(u)\right| + d_t\left(v_0, N_t^2(u)\right)} \cdot \frac{1}{n} \\
&\geq \frac{|R_t^0| - \delta_0/8}{\left|N_t^1(u)\right| + d_t\left(v_0, N_t^2(u)\right)} \cdot \frac{1}{n} \geq \frac{d_t\left(v_0, N_t^2(u)\right) - \delta_0/8 - \delta_0/8}{\left|N_t^1(u)\right| + d_t\left(v_0, N_t^2(u)\right)} \cdot \frac{1}{n} \\
&\geq \frac{3\delta_0/8 - \delta_0/8 - \delta_0/8}{\left|N_t^1(u)\right| + 3\delta_0/8} \cdot \frac{1}{n} \geq \frac{3\delta_0/8 - \delta_0/8 - \delta_0/8}{(1+1/8)\delta_0 + 3\delta_0/8} \cdot \frac{1}{n} = \frac{1}{12n}
\end{aligned}$$

Let X_1 be the number of rounds it takes for e_1 to occur. When e_1 occurs, let v_1 denote a witness for e_1. We know v_1 is in $W_{t_1}^0$ for some t_1, i.e., v_1 is strongly tied to $N_{t_1}^2(u) \cap N_{t_1}^3(u)$. If $d_t\left(v_1, N_t^2(u)\right) < 3\delta_0/8$ at any point, then $d_t(u) \geq (1+1/8)\delta_0$. Thus, in the remainder of the proof, we consider the case where $d_t\left(v_1, N_t^2(u)\right) \geq 3\delta_0/8$. Let S_t^1 (resp., W_t^1) denote the set of v_1's neighbors in $N_t^2(u)$ that are strongly tied (resp., weakly tied) to $N_t^1(u)$. If $|S_t^1| \geq \delta_0/8$, then as we did for the case $|S_t^0| \geq \delta_0/8$, we argue that in $O(n \log n)$ rounds, the degree of u is at least $(1+1/8)\delta_0$ with probability at least $1 - 1/n^2$.

Thus, in the remainder, we assume that $|S_t^1| < \delta_0/8$. Define $R_t^1 = R_{t-1}^1 \cup W_t^1$, $R_{t_1}^1 = W_{t_1}^1$. Let e_2 denote the event $\left\{u \text{ connects to a node in } R_t^0 \setminus N_t^1(u) \text{ (or } R_t^1 \setminus N_t^1(u)\text{) through } v_0 \text{(or } v_1\text{) in round } t\right\}$. By the same calculation as for v_0, we have $\Pr[e_2] \geq 1/6n$. Similarly, we can define $e_3, X_3, e_4, X_4, \ldots, e_{\delta_0/4}, X_{\delta_0/4}$, and obtain that $\Pr[e_i] \geq i/(12n)$. The total number of rounds for u to gain $\delta_0/4$ edges is bounded by $T = \sum_i X_i$. By Lemma 2, $T \leq 36n \ln n$ with probability at least $1 - 1/n^2$, completing the proof of this lemma. □

Lemma 7 (When all neighbors are weakly tied to two-hop neighbors). *There exists $T = O(n \log n)$ such that if all nodes in $N_t^1(u)$ are weakly tied to $N_t^2(u)$ for all $t < T$, then $d_T(u) \geq \min\{(1+1/8)\delta_0, n-1\}$ with probability at least $1 - 1/n^2$.*

Proof. If at any point $t < T$, $d_t(u) \geq \min\{(1+1/8)\delta_0, n-1\}$, then the claim of this lemma holds. In the remainder of this proof, we assume $d_t(u) < \min\{(1+1/8)\delta_0, n-1\}$ for all $t < T$. In the following, we first show, any node $v \in N_0^2(u)$ will have at least $\delta_0/4$ edges to $N_{T_1}^1(u)$, where $T_1 = O(n \log n)$. After that, v will connect to u in $T_2 = O(n \log n)$ rounds. Therefore, the total number of rounds used for v to connect to u is $T_3 = T_1 + T_2 = O(n \log n)$.

Node v at least connects to one node in $N_0^1(u)$. Call it w_1. Because all nodes in $N_t^1(u)$ are weakly tied to $N_t^2(u)$, we have $d_t\left(w_1, N_t^1(u)\right) \geq \delta_0 - \delta_0/2 = \delta_0/2$. If $d_t\left(w_1, N_t^1(u) \setminus N_t^1(v)\right) < \delta_0/4$, then v already has $\delta_0/4$ edges to $N_t^1(u)$. Thus, in the following we consider the case where $d_t\left(w_1, N_t^1(u) \setminus N_t^1(v)\right) \geq \delta_0/4$. Let e_1 denote the event $\left\{v \text{ connects to a node in } N_t^1(u) \setminus N_t^1(v) \text{ through } w_1\right\}$.

$$\begin{aligned}
\Pr[e_1] &= \frac{d_t\left(w_1, N_t^1(u) \setminus N_t^1(v)\right)}{d_t(w_1)} \cdot \frac{1}{d_t(w_1)} \geq \frac{d_t\left(w_1, N_t^1(u) \setminus N_t^1(v)\right)}{\left|N_t^1(u)\right| + d_t\left(w_1, N_t^2(u)\right)} \cdot \frac{1}{d_t(w_1)} \\
&\geq \frac{\delta_0/4}{(1+1/8)\delta_0 + \delta_0/2} \cdot \frac{1}{d_t(w_1)} \geq \frac{2}{13} \cdot \frac{1}{n} > \frac{1}{7n}
\end{aligned}$$

Let X_1 be the number of rounds needed for e_1 to occur. When e_1 occurs, let w_2 denote a witness for e_1. Notice w_2 is also weakly tied to $N_t^2(u)$. By similar argument, we have $d_t\left(w_2, N_t^1(u) \setminus N_t^1(v)\right) \geq \delta_0/4$. Let e_2 denote the event $\left\{v \text{ connects to a node in } N_t^1(u) \text{ through } w_1 \text{ or } w_2\right\}$. We have $\Pr[e_2] \geq 2/(7n)$. Let X_2 be the number of rounds needed for e_2 to occur. Similarly, we can define $e_3, X_3, \ldots, e_{\delta_0/4}, X_{\delta_0/4}$ and show $\Pr[e_i] \geq i/(7n)$. Set $T_1 = \sum_i X_i$, which is the bound on the number of rounds needed for v to have at least $\delta_0/4$ neighbors in $N_t^1(u)$. By Lemma 2, we know $T_2 \leq 28n \ln n$ with probability at least $1 - 1/n^3$. Now we show v will connect to u in T_2 time after this. Notice that, all w_i's are still weakly tied to $N_t^2(u)$. By Lemma 4, the probability that u connects to v through w_i in round t is at least $1/(4\delta_0^2)$. We have $w_1, w_2, \ldots, w_{\delta_0/4}$ independently executing a triangulation step each round. Consider $T_2 = 48\delta_0 \ln n$ rounds of the process. Then,

$$\Pr[u \text{ connects to } v \text{ in } T_2 \text{ rounds}] \geq 1 - \left(1 - \frac{1}{4\delta_0^2}\right)^{12\delta_0^2 \ln n} \geq 1 - \frac{1}{n^3}.$$

Combine the two steps. We have shown for any node $v \in N_0^2(u)$, it will connect to u in time $T_3 = T_1 + T_2$ with probability at least $1 - 1/n^3$. This implies in time T_3, u will connect to all nodes in $N_0^2(u)$ with probability at least $1 - \left|N_0^2(u)\right|/n^3$. Then, $N_0^2(u) \subseteq N_{T_3}^1(u), N_0^3(u) \subseteq N_{T_3}^1(u) \cup N_{T_3}^2(u), N_0^4(u) \subseteq N_{T_3}^1(u) \cup N_{T_3}^2(u) \cup N_{T_3}^3(u)$. Now we apply the above analysis twice, and obtain that in time $T = 3T_3 = O(n \log n)$, $N_0^2(u) \cup N_0^3(u) \cup N_0^4(u) \subseteq N_T^1(u)$ with probability at least $1 - \left|N_0^2(u) \cup N_0^3(u) \cup N_0^4(u)\right|/n^3 \geq 1 - 1/n^2$. By Lemma 1, we know $\left|N_0^2(u) \cup N_0^3(u) \cup N_0^4(u)\right| \geq \min\{2\delta_0, n-1\}$. Thus, we complete the proof of this lemma. □

Theorem 8 (Upper bound for triangulation process). *For any connected undirected graph, the triangulation process converges to a complete graph in $O(n \log^2 n)$ rounds with high probability.*

Proof. We first show that in $O(n \log n)$ rounds, either the graph becomes complete or the minimum degree of the graph increases by a factor of at least $1/12$. Then we apply this argument $O(\log n)$ times to complete the proof of this theorem.

For each u where $d_0(u) < \min\{(1 + 1/8)\delta_0, n-1\}$, we consider the following 2 cases. The first case is if more than $\delta_0/3$ nodes in $N_0^1(u)$ are strongly tied to $N_0^2(u)$. By Lemma 5, there exists $T = O(n \log n)$ such that if at least $\delta_0/4$ nodes in $N_t^1(u)$ are strongly tied to $N_t^2(u)$ for $t < T$, then $d_T(u) \geq (1 + 1/8)\delta_0$ with probability at least $1 - 1/n^2$. Whenever the condition is not satisfied, i.e., less than $\delta_0/4$ nodes in $N_t^1(u)$ are strongly tied to $N_t^2(u)$, it means more than $\delta_0/3 - \delta_0/4 = \delta_0/12$ strongly tied nodes became weakly tied. By the definitions of strongly tied and weakly tied, this implies $d_T(u) \geq (1 + 1/12)\delta_0$.

The second case is if less than $\delta_0/3$ nodes in $N_0^1(u)$ are strongly tied to $N_0^2(u)$. By Lemmas 6 and 7, we know that there exists $T = O(n \log n)$ such that if we remain in this case for T rounds, then $d_T(u) \geq \min\{(1 + 1/8)\delta_0, n-1\}$ with probability at least $1 - 1/n^2$. Whenever the condition is not satisfied, i.e., more than $\delta_0/3$ nodes in $N_t^1(u)$ are strongly tied to $N_t^2(u)$, we move to the analysis in the first case, and $d_T(u) \geq (1 + 1/8)\delta_0$ in $T = O(n \log n)$ time with probability at least $1 - 1/n^2$.

Combining the above 2 cases and applying a union bound, we obtain $\delta_T \geq \min\{(1 + 1/8)\delta_0, n-1\}$ in $T = O(n \log n)$ rounds with probability at least $1 - 1/n$. We now apply the above argument $O(\log n)$ times to obtain the desired $O(n \log^2 n)$ upper bound. □

3.2 Lower bound

Theorem 9 (Lower bound for triangulation process). *For any connected undirected graph G that has $k \geq 1$ edges less than the complete graph the triangulation process takes $\Omega(n \log k)$ steps to complete with probability at least $1 - O\left(e^{-k^{1/4}}\right)$.*

Proof. We first observe that during the triangulation process there is a time t when the number of missing edges is at least $m = O(\sqrt{k})$ and the minimum degree is at least $n/3$. If $k < \frac{2}{3}n$ then this is true initially and for larger k this is true at the first time t the minimum degree is large enough. The second case follows since the degree of a node (and thus also the minimum degree) can at most double in each step guaranteeing that the minimum degree is not larger than $\frac{2}{3}n$ at time t also implying that at least $\frac{n}{3} = \Omega(\sqrt{k})$ edges are still missing.

Given the bound on the minimum degree any missing edge $\{u, v\}$ is added by a fixed node w with probability at most $\frac{9}{2n^2}$. Since there are at most $n-2$ such nodes the probability that a missing edge gets added is at most $\frac{9}{2n}$. To analyze the time needed for all missing edges to be added we denote with X_i the random variable counting the number of steps needed until the ith of the m missing edges is added. We would like to analyze $\Pr[X_1 \leq T, X_2 \leq T, \ldots, X_m \leq T]$ for an appropriately chosen number of steps T. Note that the events $X_i < T$ and $X_j < T$ are not independent and indeed can be positively or negatively correlated. Nevertheless, independent of the conditioning onto any of the events $X_j < T$, we have that $\Pr[X_1 \leq T] \leq 1 - (1 - \frac{9}{2n})^T \leq 1 - \frac{1}{\sqrt{m}}$ for an appropriately chosen $T = \Omega(n \log m)$, where m is again the number of missing edges at time t. Thus,

$$\Pr[X_1 \leq T, X_2 \leq T, \ldots, X_m \leq T] =$$

$$= \Pr[X_1 \leq T | X_2 \leq T, \ldots, X_m \leq T] \cdot \Pr[X_2 \leq T | X_3, \ldots, X_m \leq T] \cdot \ldots \cdot \Pr[X_m \leq T]$$

$$\leq \left(1 - \frac{1}{\sqrt{m}}\right)^m = O\left(e^{-\sqrt{m}}\right) = O\left(e^{-k^{1/4}}\right)$$

This shows that the triangulation process takes with probability at least $1 - O\left(e^{-k^{1/4}}\right)$ at least $\Omega(n \log m) = O(n \log k)$ steps to complete. □

4 The two-hop walk: Discovery through pull

In this section, we analyze the two-hop walk process on undirected connected graphs, which is described by the following simple iteration: In each round, for each node u, we add edge (u, w) where w is drawn uniformly at random from $N_t^1(v)$, where v is drawn uniformly at random from $N_t^1(u)$. The two-hop walk yields the following pull-based resource discovery protocol. In each round, each node u contacts a random neighbor v, receives the identity of a random neighbor w of v, and sends its identity to w. The main result of this section is that the two-hop walk process transforms an arbitrary connected n-node graph to a complete graph in $O(n \log^2 n)$ rounds with high probability. We also establish an $\Omega(n \log n)$ lower bound on the two-hop walk for almost all n-node graphs.

4.1 Upper bound

As for the triangulation process, we establish the $O(n \log^2 n)$ upper bound by showing that the minimum degree of the graph increases by a constant factor (or equals $n-1$) in $O(n \log n)$ rounds with high probability. For analyzing the growth in the degree of a node u, we consider two overlapping cases. The first case is when the two-hop neighborhood of u is not too large, i.e., $|N_t^2(u)| < \delta_0/2$, and the second is when the two-hop neighborhood of u is not too small, i.e., $|N_t^2(u)| \geq \delta_0/4$. As in the analysis of the triangulation process, we also use the notions of strongly and weakly tied based on how many edges connect a node to a given set; it is more convenient to work with a different threshold. We say that a node v is ***weakly tied*** to a set of nodes S if v has less than $\delta_0/4$ edges to S (i.e. $d_t(v, S) < \delta_0/4$), and ***strongly tied*** to S if v has at least $\delta_0/4$ edges to S (i.e. $d_t(v, S) \geq \delta_0/4$).

Lemma 10 (When the two-hop neighborhood is not too large). *There exists $T = O(n \log n)$ such that either $\left|N_T^2(u)\right| \geq \delta_0/2$ or $d_T(u) \geq \min\{2\delta_0, n-1\}$ with probability at least $1 - 1/n^2$.*

Proof. By the definition of δ_0, $d_0(w) \geq \delta_0$ for all w in $N_0^1(u)$. Let X be the first round at which $|N_X^2(u)| \geq \delta_0/2$. We consider two cases. If X is at most $cn \log n$ for a constant c to be specified later, then the claim of the lemma holds. In the remainder of this proof we consider the case where X is greater than $cn \log n$; thus, for $0 \leq t \leq cn \log n$, $|N_t^2(u)| < \delta_0/2$.

Consider any node w in $N_0^1(u)$. Since $d_0(w) \geq \delta_0$ and $|N_t^2(u)| < \delta_0/2$, w has at least $\delta_0/2$ edges to nodes in $N_0^1(u)$. Fix a node v in $N_0^2(u)$. In the following, we first show that in $O(n \log n)$ rounds, v is strongly tied to the neighbors of u with probability at least $1 - 1/n^3$. Let T_1 denote the first round at which v has is strongly tied to $N_{T_1}^1(u)$, i.e., when $|N_{T_1}^1(v) \cap N_{T_1}^1(u)| \geq \delta_0/4$. We know that v has at least one neighbor, say w_1, in $N_0^1(u)$. Consider any $t < T_1$. Since v is weakly tied to $N_0^1(u)$ at time t, w_1 has at least $\delta_0/4$ neighbors in $N_0^1(u)$ which do not have an edge to v at time t. This implies

$$\Pr\left[v \text{ connects to a node in } N_0^1(u) \text{ through } w_1 \text{ in round } t\right] \geq \frac{1}{n} \cdot \frac{1}{4} = \frac{1}{4n}$$

Let e_1 denote the event $\left\{v \text{ connects to a node in } N_0^1(u)\right\}$, and X_1 be the number of rounds for e_1 to occur. When e_1 occurs, let w_2 denote a witness for e_1. We note that $w_1, w_2 \in N_0^1(u) \subseteq N_{X_1}^1(u)$. If v is weakly tied to $N_{X_1}^1(u)$, both w_1 and w_2 have at least $\delta_0/4$ neighbors in $N_{X_1}^1(u)$ that do not have an edge to v yet. Let e_2 denote the event $\left\{v \text{ connects to a node in } N_{X_1}^1(u)\right\}$, and X_2 be the number of rounds for e_2 to occur. Then $\Pr[e_2] = 2\Pr[e_1] \geq 1/2n$. Similarly, we define $e_3, X_3, \ldots, e_{\delta_0/4}, X_{\delta_0/4}$ and obtain $\Pr[e_i] \geq i/(4n)$. We now apply Lemma 2 to obtain that $X_1 + X_2 + \ldots X_{\delta_0/4}$ is at most $16n \ln n$ with probability at least $1 - 1/n^3$. Thus, with probability at least $1 - |N_0^2(u)|/n^3$, $T_1 \leq 16n \ln n$. After T_1 rounds, we obtain that for any $v \in N_0^2(u)$,

$$\Pr[u \text{ connects to } v \text{ in a single round}] \geq \frac{\delta_0/4}{2\delta_0} \cdot \frac{1}{n} = \frac{1}{8n}.$$

which implies that with probability at least $1 - 1/n^3$, u has an edge to every node in $N_0^2(u)$ in another $T_2 \leq 24n \ln n$ rounds.

Let T_3 equal $T_1 + T_2$; we set c to be at least $120 \ln 2$ so that $X > 3T_3$. We thus have $N_0^2(u) \subseteq N_{T_3}^1(u)$, $N_0^3(u) \subseteq N_{T_3}^1(u) \cup N_{T_3}^2(u)$, and $N_0^4(u) \subseteq N_{T_3}^1(u) \cup N_{T_3}^2(u) \cup N_{T_3}^3(u)$. We now repeat the above analysis again twice and obtain that at time $T = 3T_3$, $N_0^2(u) \cup N_0^3(u) \cup N_0^4(u) \subseteq N_T^1(u)$ with probability at least $1 - \left|N_0^2(u) \cup N_0^3(u) \cup N_0^4(u)\right|/n^3 \geq 1 - 1/n^2$. By Lemma 1, we have $\left|N_T^1(u)\right| \geq \min\{2\delta_0, n-1\}$, thus completing the proof of the lemma. □

Lemma 11 (When the two-hop neighborhood is not too small). *There exists $T = O(n \log n)$ such that either $\left|N_T^2(u)\right|$ is less than $\delta_0/4$ or $d_T(u)$ is at least $\min\{(1 + 1/8)\delta_0, n-1\}$, with probability at least $1 - 1/n^2$.*

Proof. Let X be the first round at which $N_X^2(u) < \delta_0/4$. We consider two cases. If X is at most $cn \log n$ for a constant c to be specified later, then the claim of the lemma holds. In the remainder of this proof we consider the case where X is greater than $cn \log n$; thus, for $0 \leq t \leq cn \log n$, $\left|N_t^2(u)\right| \geq \delta_0/4$. If $v \in N_0^2(u)$ is strongly tied to $N_0^1(u)$, then

$$\Pr[u \text{ connects to } v \text{ in a single round}] \geq \frac{d_t\left(v, N_0^1(u)\right)}{\left|N_t^1(u)\right|} \cdot \frac{1}{n} \geq \frac{\delta_0/4}{(1 + 1/8)\delta_0} \cdot \frac{1}{n} = \frac{2}{9n}$$

Thus, in $T = 13.5n \ln n$ rounds, u will add an edge to v with probability at least $1 - 1/n^3$. If there are at least $\delta_0/8$ nodes in $N_0^2(u)$ that are strongly tied to $N_0^1(u)$, then u will add edges to all these nodes in T rounds with probability at least $1 - 1/n^2$.

In the remainder of this proof, we focus on the case where the number of nodes in $N_0^2(u)$ that are strongly tied to $N_0^1(u)$ at the start of round 0 is less than $\delta_0/8$. In this case, because $\left|N_t^2(u)\right| \geq \delta_0/4$, more than $\delta_0/8$ nodes in $N_0^2(u)$ are weakly tied to $N_0^1(u)$, and, thus, have at least $3\delta_0/4$ edges to nodes in $N_0^2(u) \cup N_0^3(u)$.

In the following we show u will connect to $\delta_0/8$ nodes in $O(n \log n)$ rounds with probability at least $1 - 1/n^2$. For any round t, let W_t denote the set of nodes in $N_t^2(u)$ that are weakly tied to $N_t^1(u)$. We refer to a length-2 path from u to a node two hops away as an *out-path*. Let P_0 denote the set of out-paths to W_0. Since we have at least $\delta_0/8$ nodes in $N_0^2(u)$ that are weakly tied to $N_0^1(u)$, $|P_0|$ is at least $\delta_0/8$ at time $t = 0$. Define $e_1 = \left\{u \text{ picks an out-path in } P_0 \text{ and connects to node } v_1 \text{ in } N_0^2(u)\right\}$, and X_1 to be the number of rounds for e_1 to occur. When $0 \leq t \leq X_1$, for each $w_i \in N_t^1(u)$, let f_i be the number of edges from w_i to nodes in $N_t^1(u) \cup N_t^2(u)$, and p_i be the number of edges from w_i to nodes in $N_0^2(u)$ that are weakly tied to $N_0^1(u)$.

$$\begin{aligned}\Pr[e_1] &= \sum_i \frac{1}{d_t(u)} \cdot \frac{p_i}{f_i} \geq \sum_i \frac{1}{d_t(u)} \cdot \frac{p_i}{n-1} = \frac{\sum_i p_i}{(1+1/8)\delta_0(n-1)} \\ &= \frac{|S|}{(1+1/8)\delta_0(n-1)} \geq \frac{\delta_0/8}{(1+1/8)\delta_0(n-1)} \geq \frac{1}{9n}.\end{aligned}$$

After X_1 rounds, u will pick an out-path in P_0 and connect such a v_1. Define P_1 to be a set of out-paths from u to W_{X_1}. We now place a lower bound on $|P_1 \setminus P_0|$. Since $v_1 \in N_0^2(u)$ is added to $N_{X_1}^1(u)$, those out-paths in P_0 consisting of edges from v_1 to nodes in $N_0^1(u)$ are not in P_1. The number of out-paths we lose because of this is at most $\delta_0/4$. But v_1 also has at least $3\delta_0/4$ edges to $N_0^2(u) \cup N_0^3(u)$. The end points of these edges are in $N_{X_1}^1(u) \cup N_{X_1}^2(u)$. If more than $\delta_0/8$ of them are in $N_{X_1}^1(u)$, then $d_{X_1}(u) \geq (1+1/8)\delta_0$. Now let's consider the case that less than $\delta_0/8$ such end points are in $N_{X_1}^1(u)$. This means the number of edges from v_1 to $N_{X_1}^2(u)$ is at least $3\delta_0/4 - \delta_0/4 - \delta_0/8 = 3\delta_0/8$. Among the end points of these edges, if more than $\delta_0/8$ of them are strongly tied to $N_{X_1}^1(u)$, then the degree of u will become at least $(1+1/8)\delta_0$ in $O(n \log n)$ rounds with probability $1 - 1/n^2$ by our earlier argument. If not, we know that more than $\delta_0/4$ newly added edges are pointing to nodes that are weakly tied to $N_{X_1}^1(u)$. Thus, $|P_1 \setminus P_0|$ is by at least $\delta_0/4$. $|S| \geq 2 \cdot \delta_0/8$. Define $e_2 = \{u \text{ picks an out-path in } P_1 \text{ and connects to node } v_2\}$, and X_2 to be the number of rounds for e_2 to occur. During time $X_1 \leq t \leq X_2$, $\Pr[e_2]$ is at least $2 \cdot \frac{1}{9n}$. Similarly, we define $e_3, X_3, \ldots, e_{\delta_0/8}, X_{\delta_0/8}$ and derive $\Pr[e_i] \geq i/(9n)$. By Lemma 2, the number of rounds for $d_t(u) \geq (1+1/8)\delta_0$ is bounded by

$$T = X_1 + X_2 + \cdots + X_{\delta_0/8} \leq (2+1)9n \ln n = 27n \ln n$$

with probability at least $1 - 1/n^2$, completing the proof of this lemma. □

Theorem 12 (Upper bound for two-hop walk process). *For connected undirected graphs, the two-hop walk process completes in $O(n \log^2 n)$ rounds with high probability.*

Proof. We first show that in time $T = O(n \log n)$ time, the minimum degree of the graph increases by a factor of $1/8$, i.e., $\delta_T \geq \min\{(1+1/8)\delta_0, n-1\}$. Then we can apply this argument $O(\log n)$ times, and thus, complete the proof of this theorem.

For each u where $d_0(u) < \min\{(1+1/8)\delta_0, n-1\}$, we analyze by the following 2 cases. First, if $|N_0^2(u)| \geq \delta_0/2$, by Lemma 11 we know as long as $|N_t^2(u)| \geq \delta_0/4$ for all $t \geq 0$, $d_T(u) \geq \min\{(1+1/8)\delta_0, n-1\}$ with probability $1 - 1/n^2$ where $T = O(n \log n)$. Whenever the condition is not satisfied, we know at least $\delta_0/4$ nodes in $N_0^2(u)$ has been moved to $N_T^1(u)$, which means $d_T(u) \geq \min\{(1+1/4)\delta_0, n-1\}$.

Second, if $|N_0^2(u)| < \delta_0/2$, by Lemma 10 we know as long as $|N_t^2(u)| < \delta_0/2$ for all $t \geq 0$, $d_T(u) \geq \min\{(1+1/8)\delta_0, n-1\}$ with probability $1 - 1/n^2$ where $T = O(n \log n)$. Whenever the

condition is not satisfied, we are back to the analysis in the first case, and the minimum degree will become $\min\{(1+1/8)\delta_0, n-1\}$ with probability $1-1/n^2$.

Combine the above 2 cases, since we at most have n nodes whose degree is between δ_0 and $\min\{(1+1/8)\delta_0, n-1\}$, the minimum degree of G will become at least $\min\{(1+1/8)\delta_0, n-1\}$ in $O(n\log n)$ rounds with probability $1-1/n$.

Now we can apply the above argument $O(\log n)$ times, and have shown the two-hop walk process completes in $O(n\log^2 n)$ with high probability. □

4.2 Lower bound

Theorem 13 (Lower bound for two-hop walk process). *For any connected undirected graph G that has $k \geq 1$ edges less than the complete graph the two-hop process takes $\Omega(n\log k)$ steps to complete with probability at least $1 - O\left(e^{-k^{1/4}}\right)$.*

The proof of Theorem 13 is essentially the same as Theorem 9, and is omitted here.

5 Two-hop walk in directed graphs

In this section, we analyze the two-hop walk process in directed graphs. We say that the process terminates at time t if for every node u and every node v, G_t contains the edge (u,v) whenever u has a path to v in G_0.

Theorem 14. *On any n-node directed graph, the two-hop walk terminates in $O(n^2\log n)$ rounds with high probability. Furthermore, there exists a (weakly connected) directed graph for which the process takes $\Omega(n^2\log n)$ rounds to terminate.*

Proof. Consider any pair of nodes, u and v. Consider a shortest path from u to v $(v_0, v_1, v_2, \ldots, v_m)$, where $v_0 = u$, $v_m = v$ and $m \leq n$. Fix a time step t. Let e_i denote the event an edge is added from v_i to v_{i+2} in step t. The probability of occurrence of e_i is $\Pr[e_i] \geq 1/n^2$. All the e_i's are independent from one another.

$$\begin{aligned}
\Pr[\cup_i e_i] &\geq \sum_i \Pr[e_i] - \sum_i \sum_j \Pr[e_i \cap e_j] \\
&= \sum_i \Pr[e_i] - \sum_i \sum_j \Pr[e_i]\Pr[e_j] \\
&\geq m\frac{1}{n^2} - m(m-1)\frac{1}{n^4} \\
&\geq \frac{m}{n^2}
\end{aligned}$$

Let X_1 denote the number of steps it takes for the length of the above path to decrease by 1. It is clear that $E[X_1] \leq n^2/m$. In general, let X_i denote the number of steps it takes for the length of the above path to decrease by i. By Lemma 2, the number of steps it takes for the above path to shrink to an edge is at most $4n^2\ln n$ with probability $1/n^3$. Taking a union bound over all the edges yields the desired upper bound.

For the lower bound, consider a graph G_0 with the node set $\{1, 2, \ldots, n\}$ and the edge set

$$\{(3i,j),(3i+1,j) : 0 \leq i < n/4, 3n/4 \leq j < n\} \bigcup \{(3i,3i+1),(3i+1,3i+2) : 0 \leq i < n/4\}.$$

The only edges that need to be added by the two-hop process are the edges $(3i, 3i+2)$ for $0 \leq i < n/4$. The probability that node $3i$ adds the edge $(3i, 3i+2)$ in any round is at most $16/n^2$. The probability that edge $(3i, 3i+2)$ is not added in $(n^2\ln n)/32$ rounds is at least $1/\sqrt{n}$. Since the events associated with adding each of these edges are independent, the probability that all the $n/3$ edges are added in $(n^2\ln n)/32$ rounds is at most $(1-1/\sqrt{n})^{n/3} \leq e^{-\sqrt{n}/3}$, completing the lower bound proof. □

The lower bound in the above theorem takes advantage of the fact that the initial graph is not strongly connected. Extending the above analysis for strongly connected graphs appears to be much more difficult since the events corresponding to the addition of new edges interact in significant ways. We present an $\Omega(n^2)$ lower bound for a strongly connected graph by a careful analysis that tracks the event probabilities with time and takes dependencies into account.

Theorem 15. *There exists a strongly connected directed graph G_0 for which the expected number of rounds taken by the two-hop process is $\Omega(n^2)$.*

Proof. The graph $G_0 = (V, E)$ is depicted in Figure 3 and formally defined as $G_0 = (V, E)$ where $V = \{1, 2, \ldots, n\}$ with n being even, and

$$E = \{(i, j) : 1 \leq i, j \leq n/2\} \cup \{(i, i+1) : n/2 \leq i < n\} \cup \{(i, j) : i > j, i > n/2, i, j \in V\}.$$

We first establish an upper bound on the probability that edge $(i, i+h)$ is added by the start of round t, for

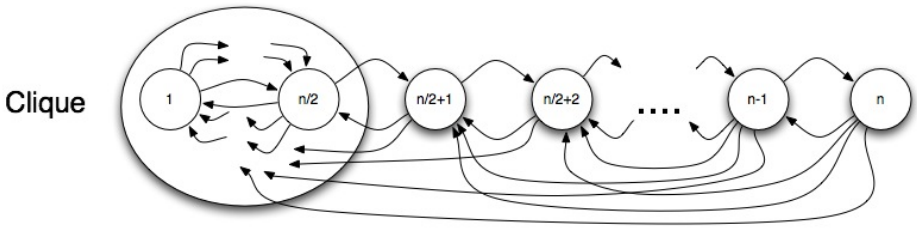

Figure 3: Lower bound example for two-hop walk process in directed graphs

given i, $1 \leq i \leq n - h$. Let $p_{h,t}$ denote this probability. The following base cases are immediate: $p_{h,0}$ is 1 for $h = 1$ and $h < 0$, and 0 otherwise. Next, the edge $(i, i+h)$ is in G_{t+1} if and only if $(i, i+h)$ is either in G_{t-1} or added in round t. In the latter case, $(i, i+h)$ is added by a two-hop walk $i \to i+k \to i+h$, where $-i < k \leq n - i$. Since the out-degree of every node is at least $n/2$, for any k the probability that i takes such a walk is at most $4/n^2$.

$$\begin{aligned} p_{h,t+1} &\leq p_{h,t} + \frac{4}{n^2} \sum_{k>-i}^{n-i} p_{k,t} p_{h-k,t} \\ &= p_{h,t} + \frac{4}{n^2} \left(\sum_{k=1}^{i-1} p_{h+k,t} + \sum_{k=1}^{h-1} p_{k,t} p_{h-k,t} + \sum_{k=h+1}^{n-i} p_{k,t} \right) \end{aligned} \tag{1}$$

We show by induction on t that

$$p_{h,t} \leq \left(\frac{\alpha t}{n^2} \right)^{h-1}, \text{ for all } t \leq \epsilon n^2 \tag{2}$$

where α and ϵ are positive constants that are specified later.

The induction base is immediate. For the induction step, we use the induction hypothesis for t and

Equation 1 and bound $p_{h,t+1}$ as follows.

$$\begin{aligned}
p_{h,t+1} &\leq \left(\frac{\alpha t}{n^2}\right)^{h-1} + \frac{4}{n^2}\left(\sum_{k=1}^{i-1}\left(\frac{\alpha t}{n^2}\right)^{h+k-1} + \sum_{k=1}^{h-1}\left(\frac{\alpha t}{n^2}\right)^{k-1}\left(\frac{\alpha t}{n^2}\right)^{h-k-1} + \sum_{k=h+1}^{n-i}\left(\frac{\alpha t}{n^2}\right)^{k-1}\right) \\
&\leq \left(\frac{\alpha t}{n^2}\right)^{h-1} + \frac{4}{n^2}\left((h-1)\left(\frac{\alpha t}{n^2}\right)^{h-2} + \left(\frac{\alpha t}{n^2}\right)^{h}\frac{2}{1-\alpha t/n^2}\right) \\
&\leq \left(\frac{\alpha t}{n^2}\right)^{h-1} + (h-1)\left(\frac{\alpha t}{n^2}\right)^{h-2}\frac{1}{n^2}\left(4 + \frac{4\epsilon^2}{(1-\alpha\epsilon)}\right) \\
&\leq \left(\frac{\alpha t}{n^2}\right)^{h-1} + (h-1)\left(\frac{\alpha t}{n^2}\right)^{h-2}\frac{\alpha}{n^2} \\
&\leq \left(\frac{\alpha(t+1)}{n^2}\right)^{h-1}.
\end{aligned}$$

(In the second inequality, we combine the first and third summations and bound them by their infinite sums. In the third inequality, we use $t \leq \epsilon n^2$. For the fourth inequality, we set α sufficiently large so that $\alpha \geq 4 + 4/(1-\alpha\epsilon)$. The final inequality follows from Taylor series expansion.)

For an integer x, let C_x denote the cut $(\{u : u \leq x\}, \{v, v > x\})$. We say that a cut C_x is *untouched* at the start of round t if the only edge in G_t crossing the cut C_x is the edge $(x, x+1)$; otherwise, we say C_x is *touched*. Let X denote the smallest integer such that C_X is untouched. We note that X is a random variable that also varies with time. Initially, $X = n/2$.

We divide the analysis into several phases, numbered from 0. A phase ends when X changes. Let X_i denote the value of X at the start of phase i; thus $X_0 = n/2$. Let T_i denote the number of rounds in phase i. A new edge is added to the cut C_{X_i} only if either X_i selects edge $(X_i, X_i + 1)$ as its first hop or a node $u < X_i$ selects $u \to X_i \to X_i + 1$. Since the degree of every node is at least $n/2$, the probability that a new edge is added to the cut C_i is at most $2/n + n(4/n^2) = 6/n$, implying that $E[T_i] \geq n/6$.

We now place a bound on X_{i+1}. Fix a round $t \leq \epsilon n^2$, and let E_x denote the event that C_x is touched by round t. We first place an upper bound on the probability of E_x for arbitrary x using Equation 2.

$$\Pr[E_x] \leq \sum_{h\geq 2} h\left(\frac{\alpha t}{n^2}\right)^{h-1} \leq \frac{\alpha t(4 - 3(\alpha t)/n^2 + (\alpha t)^2/n^4)}{n^2(1-(\alpha t)/n^2)^3},$$

for $t \leq \epsilon n^2$, where we use the inequality $\sum_{h\geq 2} h^2\delta^h = \delta(4 - 3\delta + \delta^2)/(1-\delta)^3$ for $0 < \delta < 1$. We set ϵ sufficiently small so that $(4 - 3\epsilon + \epsilon^2)/(1-\epsilon)^3 \leq 5$, implying that the above probability is at most 5ϵ.

If E_x were independent from E_y for $x \neq y$, then we can invoke a straightforward analysis using a geometric probability distribution to argue that $E[X_{i+1} - X_i]$ is at most $1/(1-5\epsilon) = O(1)$. The preceding independence does not hold, however; in fact, for $y > x$, $\Pr[E_y \bmod E_x] > \Pr[E_y]$. We show that the impact of this correlation is very small when x and y are sufficiently far apart. We consider a sequence of cuts $C_{x_1}, C_{x_2}, \ldots, C_{x_\ell}, \ldots$ where $x_0 = X_i + 2$ and $x_\ell = x_{\ell-1} + c\ell$, for a constant c chosen sufficiently

large. We bound the conditional probability of E_{x_ℓ} given $E_{x_{\ell-1}} \cap E_{x_{\ell-2}} \cdots E_{x_1}$ as follows.

$$\begin{aligned}
& \Pr[E_{x_\ell} | E_{x_{\ell-1}} \cap E_{x_{\ell-2}} \cdots E_{x_1}] \\
= \; & \frac{\Pr[E_{x_\ell} \cap E_{x_{\ell-1}} \cap E_{x_{\ell-2}} \cdots E_{x_1}]}{\Pr[E_{x_{\ell-1}} \cap E_{x_{\ell-2}} \cdots E_{x_1}]} \\
\leq \; & \frac{\Pr[E_{x_{\ell-1}} \cap E_{x_{\ell-2}} \cdots E_{x_1} \cap (C_{x_\ell} \cap (C_{x_{\ell-1}} \cup \cdots \cup C_{x_1}) = \emptyset)]}{\Pr[E_{x_{\ell-1}} \cap E_{x_{\ell-2}} \cdots E_{x_1}]} + \\
& \frac{\Pr[E_{x_{\ell-1}} \cap E_{x_{\ell-2}} \cdots E_{x_1} \cap (C_{x_\ell} \cap (C_{x_{\ell-1}} \cup \cdots \cup C_{x_1}) \neq \emptyset)}{\Pr[E_{x_{\ell-1}} \cap E_{x_{\ell-2}} \cdots E_{x_1}]} \\
\leq \; & \frac{\Pr[E_{x_{\ell-1}} \cap E_{x_{\ell-2}} \cdots E_{x_1}] \Pr[\text{a new edge is added from } (x_{\ell-1}+1, x_\ell) \text{ to } (x_\ell+1, n]]}{\Pr[E_{x_{\ell-1}} \cap E_{x_{\ell-2}} \cdots E_{x_1}]} \\
& \frac{\Pr[\text{an edge spanning at least } c\ell \text{ hops is added across } C_{x_\ell}]}{\Pr[E_{x_{\ell-1}} \cap E_{x_{\ell-2}} \cdots E_{x_1}]} \\
\leq \; & \Pr[E_{x_\ell}] + \frac{((\alpha t)/n^2)^{c\ell-1}}{(1-\alpha t/n^2)^2 (t/n^2)^\ell} \\
\leq \; & 5\epsilon + \epsilon = 6\epsilon,
\end{aligned}$$

where we set c sufficiently large in the last step. Since X_{i+1} is at most the smallest x_ℓ such that C_{x_ℓ} is untouched, we obtain that

$$E[X_{i+1} - X_i] \leq 2 + \sum_{\ell \geq 2} (6\epsilon)^\ell c\ell^2 = O(1).$$

We thus obtain that after $\epsilon' n$ phases, $E[X]$ is $O(n)$, where ϵ' is chosen sufficiently small so that $n - E[X]$ is $\Omega(n)$. Since the expected length of each phase is at least $n/6$, it follows that the expected number of rounds it takes for the two-hop process to complete is $\Omega(n^2)$ rounds.

□

6 Conclusion

We have analyzed two natural gossip-based discovery processes in networks and showed almost-tight bounds on their convergence in arbitrary networks. Our processes are motivated by the resource discovery problem in distributed networks as well as by the evolution of social networks. We would like to study variants of the processes that take into account failures associated with forming connections, the joining and leaving of nodes, or having only only a subset of nodes to participate in forming connections. We believe our techniques can be extended to analyze such situations as well. From a technical standpoint, the main problem left open by our work is to resolve the logarithmic factor gap between the upper and lower bounds. It is not hard to show that from the perspective of increasing the minimum degree by a constant factor, our analysis is tight up to constant factors. It is conceivable, however, that a sharper upper bound can be obtained by an alternative analysis that uses a "smoother" measure of progress.

References

[1] I. Abraham and D. Dolev. Asynchronous resource discovery. *Computer Networks*, 50:1616–1629, July 2006.

[2] M. Adler, E. Halperin, R. M. Karp, and V. V. Vazirani. A stochastic process on the hypercube with applications to peer-to-peer networks. In *STOC*, pages 575–584, 2003.

[3] N. Alon. Problems and results in extremal combinatorics – ii. *DISCRETE MATHEMATICS*, 2003.

[4] S. Bornholdt and H. Schuster (Editors). *Handbook of Graphs and Networks*. Wiley-VCH, 2003.

[5] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Trans. on Infor. Theory*, 52(6):2508–2530, 2006.

[6] S. Chakrabarti, A. Frieze, and J. Vera. The influence of search engines on preferential attachment. In *SODA*, 2005.

[7] J. Chen and G. Pandurangan. Optimal gossip-based aggregate computation. In *SPAA*, pages 124–133, 2010.

[8] C. Cooper and A. Frieze. Crawling on web graphs. In *STOC*, 2002.

[9] Alan Demers, Dan Greene, Carl Hauser, Wes Irish, John Larson, Scott Shenker, Howard Sturgis, Dan Swinehart, and Doug Terry. Epidemic algorithms for replicated database maintenance. In *PODC*, pages 1–12, 1987.

[10] N. B. Dimitrov and C. Greg Plaxton. Optimal cover time for a graph-based coupon collector process. In *ICALP*, pages 702–716, 2005.

[11] M. Harchol-balter, T. Leighton, and D. Lewin. Resource discovery in distributed networks. In *Symposium on Principles of Distributed Computing*, pages 229–237, 1999.

[12] R. M. Karp, C. Schindelhauer, S. Shenker, and B. Vöcking. Randomized rumor spreading. In *FOCS*, pages 565–574, 2000.

[13] D. Kempe, A. Dobra, and J. Gehrke. Gossip-based computation of aggregate information. In *FOCS*, pages 482–491, 2003.

[14] D. Kempe and J. Kleinberg. Protocols and impossibility results for gossip-based communication mechanisms. In *Proceedings of The 43rd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2002.

[15] D. Kempe, J. Kleinberg, and A. Demers. Spatial gossip and resource location protocols. In *STOC*, 2001.

[16] S. Kutten, D. Peleg, and U. Vishkin. Deterministic resource discovery in distributed networks. *SPAA*, 2001.

[17] C. Law and K. Siu. An $o(\log n)$ randomized resource discovery algorithm. In *14th International Symposium on Distributed Computing (Brief Announcement), Technical Report, Technical University of Madrid*, pages 5–8, 2000.

[18] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2004.

[19] Damon Mosk-Aoyama and Devavrat Shah. Computing separable functions via gossip. In *PODC*, pages 113–122, 2006.

[20] M. J. Newman, A. Barabasi, and D. J. Watts. *Structure and Dynamics of Networks*. Princeton University Press, 2006.

[21] F. Vega-Redondo. *Complex Social Networks*. Cambridge University Press, 2007.