

# Connectivity in Random Forests and Credit Networks

Ashish Goel <sup>\*</sup>    Sanjeev Khanna<sup>†</sup>    Sharath Raghvendra <sup>‡</sup>    Hongyang Zhang <sup>§</sup>

## Abstract

Recent work has highlighted credit networks as an effective mechanism for modeling trust in a network: agents issue their own currency and trust each other for a certain amount of each other’s currency, allowing two nodes to transact if there is a chain of sufficient residual trust between them. Under a natural model of repeated transactions, the probability that two agents can successfully transact in a credit network (i.e. the *liquidity* between these two agents) is the same as the probability that they are connected to each other in a uniformly random forest of the network.

Motivated by this connection, we define the RF-connectivity between a pair of nodes in a graph  $G$  as the probability that the two nodes belong to the same connected component in a uniformly random forest of  $G$ . Our first result is that for an arbitrary subset  $S$  of nodes in  $G$ , the average RF-connectivity between pairs of nodes in  $S$  is at least  $1 - 2/h(G_S)$ , where  $h(G_S)$  is the edge expansion of the subgraph  $G_S$  induced by  $S$ . Informally, this implies that a well-connected “community” of nodes  $S$  in a credit network will have high liquidity among themselves, regardless of the structure of the remaining network. We extend this result to show that in fact every node in  $S$  has good average RF-connectivity to other nodes in  $S$  whenever  $S$  has good edge expansion. We also show that our results are nearly tight by proving an upper bound on the liquidity of regular graphs.

For our motivating application, it is important that we relate the average RF-connectivity in  $S$  to the expansion inside  $S$  and not merely to expansion of  $G$  since we would like to assert that a well-connected community has high liquidity even if the graph as a whole is not well-connected. This naturally leads to a monotonicity conjecture: *the RF-connectivity of two nodes can not decrease when a new edge is added to  $G$* . We show that the monotonicity conjecture is equivalent to showing negative correlation between inclusion of any two edges in a random forest, a long-standing open problem. Our result about the average RF-connectivity of nodes in  $S$  may be viewed as establishing a *weak version* of the monotonicity conjecture.

**Keywords.** Uniformly random forests, Liquidity in credit networks, Edge expansion, Markov chains.

<sup>\*</sup>Department of Management Science and Engineering, Stanford University. Email: [ashishg@stanford.edu](mailto:ashishg@stanford.edu). Supported in part by the DARPA GRAPHS program, via grant FA9550-12-1-0411, and NSF grant 0904325.

<sup>†</sup>Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104. Email: [sanjeev@cis.upenn.edu](mailto:sanjeev@cis.upenn.edu). Supported in part by National Science Foundation grants CCF-1116961 and IIS-1447470.

<sup>‡</sup>Supported in part by the DARPA GRAPHS program, via grant FA9550-12-1-0411.

<sup>§</sup>Department of Computer Science, Stanford University. Email: [hongyz@stanford.edu](mailto:hongyz@stanford.edu). Supported by an Enlight Foundation Engineering Fellowship.

## 1 Overview

Given a multi-graph  $G = (V, E)$ , a subset of edges is said to be a forest of  $G$  if it does not contain any cycles. In this paper, we will study uniformly random forests of  $G$ . Unlike random spanning trees [23, 7, 1, 30], several simple questions regarding random forests remain unresolved: Does there exist a polynomial time algorithm to approximately sample a random forest (or equivalently, approximately count the number of forests) of a given graph [9, 24, 2]? Does the presence of an edge in a random forest make the presence of another edge less likely [19, 29]? And of interest to us in this paper: can we characterize the component sizes of a random forest [22]?

In addition to being fundamental graph-theoretic objects (forests are independent sets of the graphic matroid, and correspond to an important class of Tutte polynomials [32]), random forests have many applications, most notably in machine learning [6, 3]. However, the motivating application for our work is *credit networks*, which have emerged as an effective mechanism for modeling trust in transaction-oriented social networks [11, 13, 15]. In a credit network, every node (i.e. user) acts as a bank and prints its own currency. Further, a weighted edge  $(u, v)$  in the network with weight  $w$  implies that  $u$  has extended a credit line of  $w$  currency units to  $v$ , i.e.,  $u$  is willing to trust up to  $w$  units of  $v$ ’s currency, at which point this credit line is saturated. However, this saturation (in fact, any transfer of currency from  $v$  to  $u$ ) results in an edge forming in the reverse direction, from  $v$  to  $u$ , since  $u$  now has some currency from  $v$  which it can return to  $v$  in exchange for some service. Payments are routed along feasible paths; if there is no way to route a payment from a payee to a payer, the transaction fails. A central question in credit networks is *liquidity*: what fraction of desired transactions between nodes  $u$  and  $v$  actually succeed? Under a natural model of transaction rates, this question, surprisingly, is equivalent to the following question [11, 20]: what is the probability that the two nodes  $u$  and  $v$  belong to the same component in a random forest of the underlying graph? This connection led to a characterization of the liquidity of credit networks that are lines, trees, and complete graphs [11]. However, realistic social networks are very far from each of these

idealized models. Instead, there is considerable evidence that real-life social networks often contain (either overlapping, or in a core-whisker structure) sub-networks or communities that have high expansion [5, 21]. This leads us to ask the following questions:

*If  $G$  has high edge-expansion, what is the probability that two nodes  $u$  and  $v$  are in the same component in a random forest of  $G$ ? And do the same bounds hold if  $u$  and  $v$  belong to a subgraph of  $G$  with high edge-expansion, even though  $G$  itself may not have high edge-expansion?*

In the rest of the paper, we will use the term “RF-connectivity of  $u$  and  $v$ ” to refer to the probability that  $u$  and  $v$  are in the same connected component in a (uniformly) random forest.

**1.1 Our results** Our main result is that for an arbitrary subset  $S$  of nodes in  $G$ , the average RF-connectivity between pairs of nodes in  $S$  is at least  $1 - 2/h(G_S)$ , where  $h(G_S)$  is the edge expansion of the subgraph  $G_S$  induced by  $S$ . Informally, this implies that a well-connected “community” of nodes  $S$  in a credit network will have high liquidity among themselves, regardless of the structure of the remaining network. Two points about this result are worth noting. First, in our motivating application, it is important that we relate the average RF-connectivity in  $S$  to the expansion inside  $S$  and not merely to expansion of  $G$  since we would like to assert that a well-connected community has high liquidity even if the graph as a whole is not well-connected. Second, since many random graph models (such as Erdős-Rényi and preferential attachment [4]) have high edge expansion, our result automatically applies to these models. For example, above the connectivity threshold, the edge expansion of an  $n$ -node Erdős-Rényi graph is  $\Omega(\log n)$ , giving an average pairwise liquidity of  $1 - O(1/\log n)$ ; this resolves an open problem from earlier work on liquidity [11].

In order to prove our main result, we use a simple Markov Chain whose state space is all forests of  $G$ . In this Markov chain, at each step we pick an edge uniformly at random. If this edge is part of the current state, the chain transitions to the forest obtained by deleting this edge. If this edge is not part of the current forest, and adding it does not cause a cycle, the edge is added to the forest. The stationary distribution of this walk is uniform; thus, to analyze the RF-connectivity of two nodes, it suffices to estimate the probability that they are in the same component in this Markov chain. This walk is simpler than the more traditional one used to generate random forests [24] which also

allows edges to be swapped. It is plausible that the more involved walk mixes faster. However, since we are not interested in designing a sampling algorithm in this paper, we choose to use the simpler walk which is easier to analyze. We essentially show that under this walk, we are very likely to have a giant component whenever the graph  $G$  has good expansion. We then extend this analysis to the case where only  $S$  has good expansion by partitioning the set of forests into classes, with each class corresponding to the set of edges outside  $S$  that are in the forest; this extension is non-trivial, and suggests a natural monotonicity conjecture, described later.

We then establish a nearly matching upper-bound: the average RF-connectivity in a  $d$ -regular graph is at most  $1 - 1/(d + 1)$ . Since the existence of  $d$ -regular graphs with  $\Omega(d)$  edge-expansion is well known [27], this yields an upper bound of  $1 - \Omega(1/h(G_S))$  on the average RF-connectivity for this graph family, where  $S = V$ . Thus, our result is essentially tight, up to constant factors in the lower order term  $1/h(G_S)$ . We also extend our main result to show that in fact every node in  $S$  has good average RF-connectivity to other nodes in  $S$  whenever  $S$  has good edge expansion, though with a slightly weaker bound of  $1 - O(\log n)/h(G_S)$ .

As mentioned before, it is important that we relate the average RF-connectivity in  $S$  to the expansion inside  $S$  and not merely to expansion of  $G$ . This naturally leads to a monotonicity conjecture: *the RF-connectivity of two nodes can not decrease when a new edge is added to  $G$* . Our result about the average RF-connectivity of nodes in  $S$  may be viewed as establishing a weak version of the monotonicity conjecture. We show that the monotonicity conjecture is equivalent to showing negative correlation between inclusion of any two edges in a random forest, a long-standing open problem [19, 25, 17]. This is in sharp contrast to the case of random spanning trees, where negative correlation is well understood [30]. We show this equivalence via an argument that directly counts the number of forests, as opposed to the Markov Chain approach outlined earlier.

While we were motivated to study RF-connectivity because of its connections to liquidity in credit networks, it is worth pointing out that this is a fundamental question in its own right. For example, distributions of component sizes of forests of complete graphs have been shown to possess interesting phase transitions [22]. Our work reinforces the importance of several open problems relating to random forests, in particular negative correlation and efficient approximate sampling (which in our application will allow an easy estimate of the liquidity between any two nodes). Further, it would be interesting to use the new analysis tools developed in this paper to study the question of strategic formation

of credit networks: *how do nodes decide how much trust to extend to each other [10, 12]?*

**Organization:** The rest of this paper is organized as follows. In Section 2 we formalize the notion of RF-connectivity and its connection to liquidity in credit networks. We present in Section 3 our results on the relationship between RF-connectivity and expansion in a (sub)graph. In Section 4, we establish the equivalence between negative correlation property and the monotonicity conjecture.

## 2 Preliminaries

**2.1 Forests and RF-connectivity** Let  $G = (V, E)$  be an *undirected* multi-graph with  $n$  vertices and  $m$  labeled edges. If there are multiple *labeled* edges between two vertices, each of them is associated with a unique label and is a distinct element in  $E$ .

A forest  $F$  is a subset of  $E$  that does not induce any cycles and contains no multiple edges. Let  $\mathcal{F}(G)$ , or simply  $\mathcal{F}$  if the graph  $G$  is clear from the context, denote the set of all forests of  $G$ . For  $1 \leq k \leq n - 1$ , let  $\mathcal{F}_k$  denote the set of forests that contain exactly  $k$  edges. Let  $\mathcal{C}(F)$  denote the set of components in  $F$ ; we represent each component by the subset of vertices it contains, instead of edges. For each vertex  $u \in V$ , let  $C_u(F)$  denote the component that contains  $u$  in  $F$ . We use  $\mathcal{F}$  to denote a uniformly random forest of  $G$ , we define:

- $\Phi_{u,v}(G) = \Pr(u \in C_v(\mathcal{F}))$  as the RF-connectivity between any two vertices  $u$  and  $v$  in  $G$ .
- $\Phi_S(G) = \sum_{u \in S} \sum_{v \in S: v \neq u} \frac{\Phi_{u,v}(G)}{|S|(|S|-1)}$  as the *average* RF-connectivity between any two vertices of  $S$  for any  $S \subseteq V$ .
- $\Phi_{u,S}(G) = \sum_{v \in S \setminus \{u\}} \frac{\Phi_{u,v}(G)}{|S|-1}$  as the *average* RF-connectivity between any vertex  $u$  of  $S$  and the rest of vertices of  $S$ .

## 2.2 RF-connectivity and Liquidity in Credit Networks

In this section, we describe the connection between credit networks and random forests. A *credit network*  $\mathcal{G} = (\mathcal{V}, \mathcal{E}; c)$  is a directed graph over agents in  $\mathcal{V}$ . Edges in the network represent pairwise credit limits between agents. A state  $s$  in the network is simply the vector of credit capacities along all the edges in the network. An edge  $(u, v) \in \mathcal{E}$  with capacity  $c_{uv}(s) > 0$  represents a credit line extended from  $u$  to  $v$  worth  $c_{uv}(s)$  units in  $v$ 's currency. Assume that capacities are integral. Successful transactions between nodes of the network result in a change in  $s$ . We denote by  $\sigma$  the set of all states that  $\mathcal{G}$  can be in.

A transaction is specified by a tuple  $\langle u, v \rangle$ , where node  $u \in \mathcal{V}$  is the payer (buyer), node  $v \in \mathcal{V}$  is the payee (seller). Given a state  $s$ , the transaction can go through provided there is a feasible path for one unit of credit flow from  $v$  to  $u$ : that is, a path  $P$  from  $v$  to  $u$  such that each edge on the path has capacity at least one. It is assumed that all currencies are the same. If the transaction goes through, then for each edge  $(u_i, u_{i+1})$  on  $P$ , the credit capacity on each goes down by one while the credit capacity on the edge  $(u_{i+1}, u_i)$  goes up by one.<sup>1</sup>

Consider a repeated transaction model where the transaction rates are given by an  $n \times n$  matrix  $\Lambda$ . The entry corresponding to  $i$ th row and  $j$ th column in  $\Lambda$  gives the probability of  $i$  initiating a transaction with  $j$ . In this paper, we work under the assumption that the transaction rates matrix is *symmetric*. Then, these repeated transactions will induce a set of equivalence classes  $\mathcal{S}$  over  $\sigma$ , and a Markov Chain whose steady state distribution is uniform over these equivalence classes.<sup>2</sup>

**DEFINITION 2.1.** (LIQUIDITY [11]) *Let  $u$  and  $v$  be two nodes in  $\mathcal{G}$  and let  $C$  be a uniformly random equivalence class of  $\mathcal{S}$ . The steady-state transaction success probability from  $u$  to  $v$  is defined to be  $\Psi_{u,v}(\mathcal{G}) = \Pr(\langle u, v \rangle \text{ goes through in } C)$ ,*

Note that  $\mathcal{G}$  induces an undirected multi-graph  $G = (V, E)$  in a natural way: the vertices  $V$  are equal to  $\mathcal{V}$  and there are  $c_{uv}(s) + c_{vu}(s)$  labeled edges between  $u$  and  $v$  in  $E$ , where  $s$  is any state in  $\sigma$ . The edges are well-defined since the sum of credit limits of  $u$  and  $v$  to each other remain the same in any state of the network.

**PROPOSITION 2.1.**  $\Phi_{u,v}(G) = \Psi_{u,v}(\mathcal{G})$  for any two nodes  $u$  and  $v$  in  $\mathcal{V}$ .

The proof of Proposition 2.1 is left to the Appendix A. Note that Proposition 2.1 also implies  $\Phi_S(G) = \Psi_S(\mathcal{G})$  for any subset of vertices  $S \subseteq \mathcal{V}$ .

## 2.3 Uniform Sampling of a Random Forest

Consider a simple random walk  $\mathcal{M}$  on  $\mathcal{F}$ : Let  $\mathcal{F}_0$  be any forest.

$$\mathcal{F}_{i+1} = \begin{cases} \mathcal{F}_i \cup \{e\} & \text{if } \mathcal{F}_i \cup \{e\} \in \mathcal{F} \\ \mathcal{F}_i \setminus \{e\} & \text{if } e \in \mathcal{F}_i \\ \mathcal{F}_i & \text{otherwise} \end{cases}$$

<sup>1</sup>Dandekar et al. [10] characterizes when there are multiple path from  $v$  to  $u$ , the choice of  $P$  is without loss of generality, by a *path-independence property*. More details can be found in Appendix A.

<sup>2</sup>We defer a formal treatment of state space and how they induce the equivalence classes to Appendix A.

where  $e$  is a uniformly random edge from  $E$ .

**PROPOSITION 2.2.** *The stationary distribution of  $\mathcal{M}$  is uniform over  $\mathcal{F}$ .*

For some other random walks that sample a uniformly random forest, we refer the reader to [9].

### 3 Connectivity in a Random Forest and Edge-Expansion

In this section, we establish a connection between average RF-connectivity in a community and the edge expansion of the community. Recall that the *edge expansion* of an undirected graph  $G$  is defined to be

$$h(G) = \min_{S \subseteq V: 0 \leq |S| \leq \frac{n}{2}} \frac{\partial(S)}{|S|}$$

where  $\partial(S)$  is the number of edges between  $S$  and  $V \setminus S$  in  $E$ , and the minimum is over all nonempty subsets of  $V$  with at most  $\frac{n}{2}$  vertices. Let  $S$  be any subset of vertices in  $G$  and  $G_S$  be the subgraph of  $G$  induced by  $S$ . Let  $h(G_S)$  denote the edge expansion of the subgraph  $G_S$ .

Our first main result is that for any subset  $S$  of vertices in the graph, the average RF-connectivity inside the subset  $S$  is at least  $1 - \frac{2}{h(G_S)}$ . Thus vertices in a community  $S$  that are well-connected to each other (i.e. have high edge-expansion in the subgraph induced by them) have high average connectivity among themselves, regardless of the structure of the remaining graph. We also show that this bound is essentially tight by establishing that average RF-connectivity in any  $d$ -regular graph is at most  $1 - 1/(d+1)$ . Our second main result is that not only is the average RF-connectivity large in a well-connected community  $S$ , but every vertex in any such community  $S$  also has high average connectivity to other nodes in  $S$ . Specifically, we will show that each vertex in a community  $S$  has average RF-connectivity at least  $1 - \frac{\ln n + 2}{h(G_S) + 1}$  inside  $S$ .

#### 3.1 Average RF-connectivity in a Community

**THEOREM 3.1.** *Given a multi-graph  $G = (V, E)$  and a subset of vertices  $S \subseteq V$ , the average RF-connectivity between pairs of nodes in  $S$  is at least  $1 - \frac{2}{h(G_S)}$ , where  $h(G_S)$  is the edge expansion of its induced subgraph  $G_S$ .*

The proof proceeds in two stages. First, we partition  $\mathcal{F}$  into classes according to their set of edges outside  $G_S$  in Lemma 3.1. Secondly, within each class of forests, we show that the average RF-connectivity is at least  $1 - \frac{2}{h(G_S)}$  in Lemma 3.3, obtaining the desired result.

We start by defining some additional notations. Let  $E_S$  be the set of edges in  $G_S$ , and let  $\overline{E_S} = E \setminus E_S$  be the set of edges outside  $G_S$ . We partition  $\mathcal{F}$  by its restriction to  $\overline{E_S}$ . We say that a set of edges  $E'$  in  $\overline{E_S}$  is *feasible* if there exists a forest  $F \in \mathcal{F}$  such that  $E' = F \cap \overline{E_S}$ . Then for any feasible set  $E'$ , we define  $\mathcal{F}_{E'}$  to be the set of forests such that  $F \cap \overline{E_S} = E'$  for any  $F \in \mathcal{F}_{E'}$ .

Consider the average connectivity of forests in  $\mathcal{F}_{E'}$ . This is captured by the RF-connectivity of a vertex-weighted multi-graph  $G(E')$ . To define this graph, note that the connectivity created by edges of  $E'$  naturally induces an equivalence relation  $\sim$  on  $S$ . For two vertices  $u, v$  in  $S$ ,  $u \sim v$  if and only if  $u$  are connected to  $v$  by  $E'$ . Let  $V^*$  denote the equivalence classes of  $\sim$  and let  $\beta : V' \rightarrow \mathbb{N}^3$  denote the number of elements in each equivalence class. Secondly, let  $I = \{(v, v', \kappa)\}$  denote the labeled edges in  $G_S$ . we define  $E^* = \{([v], [v'], \kappa)\}$ , where  $[ ]$  is the equivalence class of a vertex under  $\sim$ . Let  $G(E') = (V^*, E^*)$ .

We define  $\phi(F) = \sum_{T \in \mathcal{C}(F)} \binom{\beta(T)}{2} / \binom{\beta(V^*)}{2}$  to capture pairwise connectivity of  $F$  in this weighted graph  $G(E')$ . Lemma 3.1 shows that  $\Phi_S(G)$  can be redefined in terms of  $\phi(F)$ .

**LEMMA 3.1.** *Let  $\mathcal{E}$  be the set of feasible subsets of  $\overline{E_S}$ . Then  $\Phi_S(G) = \sum_{E' \in \mathcal{E}} \mathbb{E}[\phi(\mathcal{F})] \cdot \Pr(\mathcal{F} \cap \overline{E_S} = E')$ , where  $\mathcal{F}$  is a uniformly random forest of  $G(E')$  and  $\mathcal{F}$  is a uniformly random forest of  $G$ .*

*Proof.* Note that

$$\Phi_S(G) = \sum_{E' \in \mathcal{E}} \Pr(\mathcal{F} \cap \overline{E_S} = E') \times \left( \sum_{u \in V} \sum_{v \in V: v \neq u} \frac{\Pr(u \in C_v(\mathcal{F}) \mid \mathcal{F} \cap \overline{E_S} = E')}{|S|(|S| - 1)} \right).$$

It suffices to show that the above sum over  $u$  and  $v$  is equal to  $\mathbb{E}[\phi(\mathcal{F})]$ . We define a mapping  $\Gamma : \mathcal{F}_{E'} \rightarrow \mathcal{F}(G(E'))$  as  $\Gamma(F) = \{([v], [v'], \kappa) : \forall (v, v', \kappa) \in F \setminus E'\}$ , and prove that the mapping  $\Gamma$  is bijective:

- By the definition of  $G(E')$ , it is clear that  $\Gamma(F)$  is in  $\mathcal{F}(G(E'))$ , hence the mapping is valid.
- Two different forests from  $\mathcal{F}_{E'}$  map to different forests in  $\mathcal{F}(G(E'))$ , because they differ over the set of edges  $\overline{E_S}$ .

<sup>3</sup>Unless specified otherwise, from here on, the vertex weight function is only assumed to take positive integer values.

Note that the mapping preserves forest connectivity. That is,

$$\sum_{T \in \mathcal{C}(F)} \frac{\binom{|T \cap S|}{2}}{\binom{|S|}{2}} = \phi(\Gamma(F)),$$

since any two vertices  $v$  and  $v'$  in  $S$  are connected in  $F$  if and only if  $[v] = [v']$  or  $[v]$  are connected to  $[v']$  in  $\Gamma(F)$ . Therefore,

$$\mathbb{E} \left[ \sum_{T \in \mathcal{C}(\mathcal{F}'')} \frac{\binom{|T \cap S|}{2}}{\binom{|S|}{2}} \right] = \mathbb{E} [\phi(\Gamma(\mathcal{F}''))]$$

where  $\mathcal{F}''$  is a uniform sample from  $\mathcal{F}_{E'}$ .

It is clear that

$$\begin{aligned} & \sum_{u \in S} \sum_{v \in S: v \neq u} \frac{\Pr(u \in C_v(F) \mid F \cap \bar{E}_S = E')}{|S|(|S|-1)} \\ &= \mathbb{E} \left[ \sum_{T \in \mathcal{C}(\mathcal{F}'')} \frac{\binom{|T \cap S|}{2}}{\binom{|S|}{2}} \right] = \mathbb{E} [\phi(\Gamma(\mathcal{F}''))]. \end{aligned}$$

Finally, we observe that  $\mathbb{E} [\phi(\Gamma(\mathcal{F}''))] = \mathbb{E} [\phi(\mathcal{F})]$  since  $\Gamma(\cdot)$  is a one-to-one mapping; thus completing the proof.

We now concentrate on bounding the average RF-connectivity of  $G(E')$  for each  $E' \in \mathcal{E}$ . The notion of edge expansion function can be naturally extended to graphs with vertex weights  $\beta(\cdot)$ : we define it to be the minimum of the quantity  $\partial(S)/\beta(S)$ , taken over all sets  $S$  for which  $\beta(S)$  is at most half of the total weight. We will use the following easy but important monotonicity property of expansion.

**LEMMA 3.2.** *The edge expansion function is monotone under vertex contraction, that is,  $h(G(E')) \geq h(G_S)$  for any  $E' \in \mathcal{E}$ .*

*Proof.* It suffices to consider  $G(E')$  obtained by contracting a pair of vertices  $v$  and  $v'$  in  $G$  to  $v''$ . Let  $\partial'(\cdot)$  denote the number of edges between any subset of vertices in  $G(E')$  and its complement. Let  $T$  be any subset of vertices of  $G(E')$  such that  $\beta(T) \leq \frac{|S|}{2}$ . If  $T$  does not contain  $v''$ , then it's clear that

$$\frac{\partial'(T)}{\beta(T)} = \frac{\partial(T)}{\beta(T)} \geq h(G_S).$$

Otherwise,  $T$  contains  $v''$ ; let  $T' = T \cup \{v, v'\} \setminus \{v''\}$ . Then

$$\frac{\partial'(T)}{\beta(T)} = \frac{\partial'(T')}{\beta(T')} \geq h(G_S),$$

since  $\beta(u'') = \beta(u) + \beta(u')$  and  $\partial'(T') = \partial(T)$ . Hence edge expansion of  $G(E')$  is at least  $h(G_S)$ .

We are now ready to prove the following key lemma.

**LEMMA 3.3.** *Let  $G^* = (V^*, E^*; \beta(\cdot))$  be a multi-graph with vertex weight  $\beta(\cdot)$  and edge expansion  $h(G^*)$ . Then, for a uniformly random forest  $\mathcal{F}$  of  $G^*$ , we have  $\mathbb{E}[\phi(\mathcal{F})] \geq 1 - \frac{2}{h(G^*)}$ .*

We will prove Lemma 3.3 by analyzing the random walk  $\mathcal{M}$  introduced in 2.3. Let  $n$  be the number of vertices and  $m$  be the number of edges in  $G^*$ . Let  $\mathcal{F}^*$  denote the set of forests in  $G^*$ . We define  $\alpha(F) = \max_{T \in \mathcal{C}(F)} \beta(T)$  as the maximum weight of any tree in the forest  $F$ . Then

**LEMMA 3.4.** *For  $1 \leq k \leq n - 2$ ,*

$$\sum_{F \in \mathcal{F}_k^*} \min\left(\frac{\beta(V^*)}{2}, \beta(V^*) - \alpha(F)\right) \cdot h(G^*) \leq |\mathcal{F}_{k+1}^*| \cdot (k+1)$$

*Proof.* Consider the overall probability that any forest  $F$  with  $k+1$  edges moves to any forest  $F'$  with  $k$  edges in the random walk: it happens when any edge in  $F$  is deleted after a transition. Since there are  $k+1$  such edges, the probability of this event happening at state  $F$  is  $\frac{k+1}{m}$ . Therefore,

$$\sum_{F \in \mathcal{F}_{k+1}^*} \sum_{F' \in \mathcal{F}_k^*} \frac{P_{F,F'}}{|\mathcal{F}^*|} = \sum_{F \in \mathcal{F}_{k+1}^*} \frac{k+1}{m \cdot |\mathcal{F}^*|}$$

where  $P_{F,F'}$  is the probability  $F$  moves to  $F'$  in  $\mathcal{M}$  and  $1/|\mathcal{F}^*|$  is the stationary probability of a uniformly random forest.

Next consider the overall probability that any forest  $F$  with  $k$  edges moves to any forest  $F'$  with  $k+1$  edges: it happens when any edge between two trees of  $F$  is added after a transition in the random walk. For the number of such edges, consider two cases:

- if  $\alpha(F) \leq \frac{\beta(V^*)}{2}$ , then for each component  $C$  in  $\mathcal{C}(F)$ , the number of edges from  $C$  to the other components is at least  $h(G^*)$  times the number of vertices in  $C$ . Therefore, the total amount of edges between any two trees is at least  $h(G^*) \cdot \beta(V^*)/2$ ;
- if  $\alpha(F) > \frac{\beta(V^*)}{2}$ , then it will be at least  $(\beta(V^*) - \alpha(F))h(G^*)$ .

In summary, the total number of edges between any two trees in  $F$  is at least  $x = \min(\beta(V^*)/2, \beta(V^*) - \alpha(F)) \cdot h(G^*)$ . Hence with probability at least  $x/m$ , two trees merge together after one move from state  $F$ :

$$\begin{aligned} & \sum_{F \in \mathcal{F}_k^*} \sum_{F' \in \mathcal{F}_{k+1}^*} \frac{P_{F,F'}}{|\mathcal{F}^*|} \geq \\ & \sum_{F \in \mathcal{F}_k^*} \frac{\min\left(\frac{\beta(V^*)}{2}, \beta(V^*) - \alpha(F)\right) \cdot h(G^*)}{m \cdot |\mathcal{F}^*|} \end{aligned}$$

Since the transition matrix is symmetric ( $P_{F',F} = P_{F,F'}$ ), the Lemma is proved.

To bound the RF-connectivity, we only consider the maximum weighted component of a forest. Such an approximation is able to capture most of the connected pairs, as we will see in the following analysis. This is because we expect that there will be a giant component of size  $n - \Omega(n/h(G^*))$  in a uniformly random forest. Hence it does not lose too much to ignore the rest of components in the forest.

*Proof of Lemma 3.3:* Note that

$$\begin{aligned}\phi(F) &\geq \frac{(\alpha(F))^2 - \beta(V^*)}{(\beta(V^*))^2 - \beta(V^*)} \\ &= 1 - \frac{(\beta(V^*) - \alpha(F))(\beta(V^*) + \alpha(F))}{(\beta(V^*))^2 - \beta(V^*)} \\ &\geq 1 - \frac{2(\beta(V^*) - \alpha(F))}{\beta(V^*) - 1}\end{aligned}$$

On the other hand,  $\phi(F) \geq 0 \geq 1 - \frac{\beta(V^*)}{\beta(V^*) - 1}$ . Combined together,

$$\phi(F) \geq 1 - \frac{2 \min(\frac{\beta(V^*)}{2}, \beta(V^*) - \alpha(F))}{\beta(V^*) - 1}.$$

With this inequality,

$$\begin{aligned}&\sum_{F \in \mathcal{F}'_k} \frac{\phi(F)}{|\mathcal{F}'_k|} \\ &\geq \sum_{F \in \mathcal{F}'_k} \frac{1}{|\mathcal{F}'_k|} - \sum_{F \in \mathcal{F}'_k} \frac{2 \min(\frac{\beta(V^*)}{2}, \beta(V^*) - \alpha(F))}{(\beta(V^*) - 1) |\mathcal{F}'_k|} \\ &\geq \sum_{F \in \mathcal{F}'_k} \frac{1}{|\mathcal{F}'_k|} - \sum_{F \in \mathcal{F}'_{k+1}} \frac{2(k+1)}{h(G^*)(\beta(V^*) - 1) |\mathcal{F}'_k|} \\ &\geq \sum_{F \in \mathcal{F}'_k} \frac{1}{|\mathcal{F}'_k|} - \sum_{F \in \mathcal{F}'_{k+1}} \frac{2}{h(G^*) |\mathcal{F}'_k|}\end{aligned}$$

where we use Lemma 3.4 in the second inequality and  $\beta(V^*) \geq k + 2$  in the third inequality. By summing up this inequality from  $k = 0$  to  $n - 2$ , and taking into account that the right hand term is zero for  $k = n - 1$ ,

$$\begin{aligned}\mathbb{E}[\phi(\mathcal{F})] &= \sum_{k=0}^{n-1} \sum_{F \in \mathcal{F}'_k} \frac{\phi(F)}{|\mathcal{F}'_k|} \\ &\geq \sum_{k=0}^{n-1} \sum_{F \in \mathcal{F}'_k} \frac{1}{|\mathcal{F}'_k|} - \sum_{k=1}^{n-1} \sum_{F \in \mathcal{F}'_k} \frac{2}{h(G^*) \cdot |\mathcal{F}'_k|} \\ &\geq 1 - \frac{2}{h(G^*)}\end{aligned}$$

*Proof of Theorem 3.1:* By Lemma 3.3, for each  $E^* \in \mathcal{E}$  and a uniformly random forest  $\mathcal{F}$  of  $G(E^*)$ , we know that  $\mathbb{E}[\phi(\mathcal{F})] \geq 1 - 2/h(G(E^*))$ . Since Lemma 3.2 implies  $h(G(E^*)) \geq h(G_S)$ , therefore  $\mathbb{E}[\phi(\mathcal{F})] \geq 1 - 2/h(G_S)$ . This inequality, combined with Lemma 3.1, implies that  $\Phi_S(G) \geq 1 - 2/h(G_S)$ .

We note that the bound shown in Lemma 3.3 is asymptotically tight in the following sense. There exist  $d$ -regular graphs with edge expansion  $\Omega(d)$  (take a random  $d$ -regular graph, for instance). Lemma 3.3 asserts that average connectivity in such graphs is  $1 - \Theta(1/d)$ . On the other hand, the lemma below shows that in a  $d$ -regular graph, the average connectivity is bounded by  $1 - 1/(d + 1)$ .

**LEMMA 3.5.** *The average RF-connectivity in any  $d$ -regular graph  $G = (V, E)$  is at most  $d/(d + 1)$ .*

*Proof.* Let there be  $n$  vertices in the graph and each vertex's weight is one. We will prove a stronger result: the average RF-connectivity of any vertex  $u$  is at most  $d/(d + 1)$ .

Group  $\mathcal{F}$  by their restrictions to  $G \setminus \{u\}$ . More specifically, for any forest  $F \in \mathcal{F}(G \setminus \{u\})$ , let  $\psi(F) \subseteq \mathcal{F}$  denote the set of forests whose restriction to  $G \setminus \{u\}$  is equal to  $F$ . We claim that for a uniform sample  $\mathcal{F}$  from  $\psi(F)$ ,  $\mathbb{E}[|C_u(\mathcal{F})|] \leq (n - 1)(d/(d + 1))$ .

Let  $d_u(T)$  denote the number of edges between  $u$  and any subset  $T$  of  $V \setminus \{u\}$ . Then

$$\begin{aligned}\mathbb{E}[|C_u(\mathcal{F})|] &= \mathbb{E}\left[\sum_{T \in \mathcal{C}(\mathcal{F})} |T| \cdot \left(1 - \frac{1}{d_u(T) + 1}\right)\right] \\ &\leq \mathbb{E}\left[\sum_{T \in \mathcal{C}(\mathcal{F})} |T| \cdot \left(1 - \frac{1}{d + 1}\right)\right] = (n - 1) \cdot \frac{d}{d + 1}.\end{aligned}$$

Finally, by linearity of expectations, for a uniformly random forest  $\mathcal{F}'$  of  $G$ ,  $\mathbb{E}[|C_u(\mathcal{F}')|] \leq (n - 1) \cdot d/(d + 1)$ . Hence the average RF-connectivity of  $u$  is at most  $d/(d + 1)$ .

As another remark, the idea of Lemma 3.4 also leads to some other interesting consequences regarding uniformly random forests. Without repeating similar arguments, we simply list the results here.

**THEOREM 3.2.** *Given a graph  $G = (V, E)$  with  $n$  vertices and a uniformly random forest  $\mathcal{F}$  of  $G$ , then*

1.  $\Pr(\alpha(\mathcal{F}) \leq \frac{n}{2}) \leq \frac{2}{h(G)}$ ;
2.  $\mathbb{E}[\alpha(\mathcal{F})] \geq n(1 - \frac{2}{h(G)})$ , which also implies that the expected number of components is at most  $\frac{2n}{h(G)} + 1$ .

### 3.2 Average RF-connectivity of Any Vertex in a Community

We continue this section by establishing that not only average RF-connectivity is large when a community has good expansion but in fact every node in the community has high RF-connectivity (albeit slightly weaker than Theorem 3.1). To this end we will use more information from structures of a tree.

**THEOREM 3.3.** *Given a multi-graph  $G = (V, E)$ , a subset of vertices  $S \subseteq V$ , and a vertex  $u \in S$ , the average RF-connectivity between  $u$  and the rest of nodes in  $S$  is at least  $1 - \frac{\ln n + 2}{h(G_S) + 1}$ , where  $h(G_S)$  is the edge expansion of its induced subgraph  $G_S$ .*

The proof again proceeds in two stages. We first classify the set of forests based on how edges in  $E_S$  help connect pairs of vertices in  $S$ . Then we bound the RF-connectivity of  $u$  for each class of forests, by examining the random walk  $\mathcal{M}$  and tree structures.

We will keep using notations introduced at the beginning Section 3.1. We overload the function  $\phi(\cdot)$  and define  $\phi_u(F) = \beta(C_u(F))/\beta(V^*)$ .

**LEMMA 3.6.** *Let  $\mathcal{E}$  be the set of feasible subsets of  $\overline{E_S}$ . Then  $\Phi_{u,S}(G) = \sum_{E' \in \mathcal{E}} \mathbb{E}[\phi_u(\mathcal{F})] \cdot |\mathcal{F}_{E'}|/|\mathcal{F}|$ , where  $\mathcal{F}$  be a uniformly random forest of  $G(E')$ .*

*Proof.* Follows from the proof of Lemma 3.1.

**LEMMA 3.7.** *Let  $G^* = (V^*, E^*; \beta)$  be a vertex-weighted multi-graph with edge expansion  $h(G^*)$ . Then, for a uniformly random forest  $\mathcal{F}$  of  $G^*$ ,  $\mathbb{E}[\phi_u(\mathcal{F})] \geq 1 - \frac{\ln n + 2}{h(G^*) + 1}$ .*

The proof of Lemma 3.7 turns out to be more intricate than Lemma 3.3. When an edge is removed from the component of  $u$ , it's not impossible that more than half of the component get disconnected with  $u$ . However, it is a simple fact of spanning trees that there are at most  $k$  edge whose removal will reduce the component size of  $u$  to below  $k$ , for  $k = 1$  up to the size of this tree. With this information, we first bound the probability that the weight of  $C_u(\mathcal{F})$  is at most  $\beta(V^*)/2$ ;

**LEMMA 3.8.**  $\Pr(\beta(C_u(\mathcal{F})) \leq \frac{\beta(V^*)}{2}) \leq \frac{\ln n + 1}{h(G^*)}$ .

The following two facts will be useful in the proof of Lemma 3.8.

**FACT 3.1.** *Let  $\{x_j\}$  and  $\{y_j\}$  be two sequences of real numbers with length  $k$ . Then, subject to the set of constraints*

$$\sum_{j=1}^i j x_j \leq \sum_{j=1}^i y_j$$

for any  $i = 1, \dots, k$ , the sum of  $\{x_j\}$  is at most  $\sum_{j=1}^k y_j/j$ .

*Proof.* Multiply the  $k$ -th constraint by  $1/k$  on both sides and the  $i$ -th constraint by  $1/(i+1)$  as well, for all  $i = 1, \dots, k-1$ . Then, if we sum up the  $k$  inequalities together, it will give us the desired inequality.

It remains to verify if the coefficient of each variable equals the coefficient of the desired inequality. For  $x_k$  and  $y_k$ , it's true. For  $x_j$ , where  $j < k$ , its coefficient is given by  $\frac{j}{k} + \sum_{i=j}^{k-1} \frac{j}{i(i+1)} = 1$ . For  $y_j$ , its coefficient is given by  $\frac{1}{k} + \sum_{i=j}^{k-1} \frac{1}{i(i+1)} = \frac{1}{j}$ .

**FACT 3.2.** *Let  $T = (V, E; \beta(\cdot))$  be a vertex weighted tree of  $n$  vertices and let  $u$  be any vertex of the tree. An edge  $e$  is called  $k$ -bad if  $\beta(C_u(T \setminus \{e\})) \leq k$ , where  $1 \leq k \leq n-1$ . Then there are at most  $k$   $k$ -bad edges in  $T$ .*

*Proof.* We introduce some additional notation first. Let  $\lambda(T, k)$  denote the number of  $k$ -bad edges in  $T$ . Assume that  $T$  is a rooted tree at  $u$ , wlog. If  $v$  is a neighbor of  $u$ , then all of  $v$ 's descendants plus the root node (and the edges connecting these nodes) are called a *branch* of  $T$ . It's clear that a branch is still a rooted tree at  $u$ .

We will prove by an induction on  $k$ . The base case is easily verified. Assume that the fact holds for any cases when  $j < k$ . If  $u$  has one child, then  $\lambda(T, k) = 1 + \lambda(T', k-1) \leq k$ , where  $T'$  is the minor of  $T$  obtained by contracting  $u$  and  $u$ 's neighbor. Otherwise if the root has at least two children, then there exists a way to divide  $T$  into two trees  $T'$  and  $T''$  rooted at  $u$ , each including at least one branch of  $u$ . Let  $w'$  denote the total vertex-weights of  $T'$ . Then the total weights of  $T''$  is  $w'' = \beta(V) - w' + \beta(u)$ .

Consider the number of  $k$ -bad edges in  $T''$ : it's not zero only if  $w' \leq k$ , in which case there are at most  $\lambda(T'', k - w' + 1) \leq k - w' + 1$  (by induction hypothesis)  $k$ -bad edges. Similarly, the number of  $k$ -bad edges in  $T'$  is not zero only if  $w'' \leq k$ , and can be bounded by  $\lambda(T', k - w'' + 1) \leq k - w'' + 1$  (again by induction hypothesis). Summing up these two arguments, it becomes clear that  $\lambda(T, k) \leq \max(0, k - w' + 1, k - w'' + 1, 2k - w' - w'' + 2) \leq k$ , therefore the induction hypothesis is verified for the case of  $k$ .

*Proof of Lemma 3.8:* Let  $n$  be the number of vertices and  $m$  be the number of edges in  $G^*$ . Let  $\mathcal{F}^*$  denote the set of forests in  $G^*$ . The lemma follows from summing up the following inequality over  $k = 1, \dots, n-2$ :

$$\Pr(\mathcal{F} \in \mathcal{F}_k^* \wedge \frac{\beta(C_u(\mathcal{F}))}{\beta(V^*)} \leq \frac{1}{2}) \leq \frac{\ln n + 1}{h(G^*)} \cdot \Pr(\mathcal{F} \in \mathcal{F}_{k+1}^*) \quad (3.1)$$

Now we focus on proving inequality (3.1) for each  $k$ . For  $j = 1, \dots, \lfloor \frac{\beta(V^*)}{2} \rfloor$ , we define  $\mathcal{H}(j) \subseteq \mathcal{F}_k^*$  as the family of forests such that  $\beta(C_u(F)) = j$  for each  $F \in \mathcal{H}(j)$ . We then use the notation  $\mathcal{H}(\leq i)$  to represent the union of  $\mathcal{H}(j)$ , from  $j = 1$  up to  $i$ . It is clear that the LHS of inequality (3.1) is equal to  $\Pr(\mathcal{F} \in \mathcal{H}(\leq \lfloor \frac{\beta(V^*)}{2} \rfloor))$ . We'll find a set of constraints that govern these sets. Let  $i$  be any integer between 1 and  $\lfloor \frac{\beta(V^*)}{2} \rfloor$ .

Consider the event in  $\mathcal{M}$  when an edge in  $u$ 's component is deleted in a forest  $F$  with  $k+1$  edges, and then moves to  $F'$  such that  $\beta(C_u(F')) \leq i$ . By Fact 3.2, there're at most  $\min(|C_u(F)| - 1, i)$  such edges. By aggregating all such events, we get

$$\begin{aligned}
& \sum_{F \in \mathcal{F}_{k+1}^*} \sum_{F' \in \mathcal{H}(\leq i)} \Pr(\mathcal{F} = F) P_{F, F'} \\
& \leq \sum_{F \in \mathcal{F}_{k+1}^*} \Pr(\mathcal{F} = F) \cdot \frac{\min(|C_u(F)| - 1, i)}{m} \implies \\
& \sum_{F \in \mathcal{F}_{k+1}^*} \sum_{F' \in \mathcal{H}(\leq i)} P_{F, F'} \\
(3.2) \quad & \leq \sum_{F \in \mathcal{F}_{k+1}^*} \frac{\min(|C_u(F)| - 1, i)}{m}
\end{aligned}$$

since  $\Pr(\mathcal{F} = F)$  is the same for any forest  $F$  in  $G^*$ .

From the perspective of a forest  $F'$ , which has  $k$  edges and such that  $\beta(C_u(F')) \leq i$ , there are at least  $\beta(C_u(F'))h(G^*)$  number of ways to add an edge into  $F'$  in the next random walk. Again by aggregating all such events, we get

$$\begin{aligned}
& \sum_{F' \in \mathcal{H}(\leq i)} \frac{\beta(C_u(F'))h(G^*)}{m} \\
& \leq \sum_{F' \in \mathcal{H}(\leq i)} \sum_{F \in \mathcal{F}_{k+1}^*} P_{F', F} \implies \\
& \sum_{F' \in \mathcal{H}(\leq i)} \beta(C_u(F'))h(G^*) \\
(3.3) \quad & \leq \sum_{F \in \mathcal{F}_{k+1}^*} \min(|C_u(F)| - 1, i)
\end{aligned}$$

by applying inequality 3.2 and then note that  $P_{F, F'} = P_{F', F}$  for any pair of forests  $F$  and  $F'$ .

Now for  $j = 1, \dots, \lfloor \frac{\beta(V^*)}{2} \rfloor$ , we define  $\mathcal{H}(j)$  as the family of forests in  $\mathcal{F}_{k+1}^*$  such that  $|C_u(F)| > j$  for each  $F \in \mathcal{H}(j)$ . Then note that

$$\sum_{F \in \mathcal{H}(\leq i)} \frac{\beta(C_u(F))}{|\mathcal{F}'|} = \sum_{j=1}^i \Pr(\mathcal{F} \in \mathcal{H}(j)) \cdot j$$

and

$$\sum_{F \in \mathcal{F}_{k+1}^*} \frac{\min(|C_u(F)| - 1, i)}{|\mathcal{F}'|} = \sum_{j=1}^i \Pr(\mathcal{F} \in \mathcal{H}(j))$$

by the definition of  $\mathcal{H}(\cdot)$ .

Here we could use Fact 3.1. Let

$$x_j = \Pr(\mathcal{F} \in \mathcal{H}(j)) \cdot h(G^*) \text{ and } y_j = \Pr(\mathcal{F} \in \mathcal{H}(j))$$

note that  $y_j = 0$  when  $j \geq n$ . It's clear that inequality (3.3) shows that the two sequences  $\{x_j\}$  and  $\{y_j\}$  satisfy the set of constraints

$$\sum_{j=1}^i j x_j \leq \sum_{j=1}^i y_j,$$

for  $i = 1, \dots, \lfloor \frac{\beta(V^*)}{2} \rfloor$ . Therefore

$$\begin{aligned}
& \sum_{j=1}^{\lfloor \frac{\beta(V^*)}{2} \rfloor} \Pr(\mathcal{F} \in \mathcal{H}(j)) h(G^*) \\
& \leq \sum_{j=1}^{\lfloor \frac{\beta(V^*)}{2} \rfloor} \frac{\Pr(\mathcal{F} \in \mathcal{H}(j))}{j} \\
& = \sum_{i=2}^n \Pr(\mathcal{F} \in \mathcal{F}_{k+1} \wedge |C_u(\mathcal{F})| = i) \\
& \quad \cdot \left( \sum_{j=1}^{\min(i-1, \lfloor \frac{\beta(V^*)}{2} \rfloor)} \frac{1}{j} \right) \\
& \leq \sum_{i=2}^n \Pr(\mathcal{F} \in \mathcal{F}_{k+1} \wedge |C_u(\mathcal{F})| = i) (1 + \ln n) \\
& \leq \Pr(\mathcal{F} \in \mathcal{F}_{k+1}) (1 + \ln n)
\end{aligned}$$

Hence (3.1) is proved.

*Proof of Lemma 3.7:* Let  $\partial_u(F)$  denote the amount of edges between  $C_u(F)$  and the rest of nodes in  $G^*$ . Note that the expected number of transitions that adds one edge to  $\mathcal{F}$  is equal to the number of transitions that deletes one edge from  $\mathcal{F}$ . This is because the Markov chain  $\mathcal{M}$  is symmetric. Therefore

$$\begin{aligned}
& \mathbb{E}[|C_u(\mathcal{F})| - 1] = \mathbb{E}[\partial_u(\mathcal{F})] \\
& \geq \mathbb{E}[\min(\beta(C_u(\mathcal{F})), \beta(V^*) - \beta(C_u(\mathcal{F}))) \cdot h(G^*)].
\end{aligned}$$

After rearranging the inequality, we get

$$\mathbb{E}[\phi_u(\mathcal{F})] \geq 1 - \frac{\Pr(\frac{\beta(C_u(\mathcal{F}))}{\beta(V^*)} \leq \frac{1}{2}) \cdot h(G^*) + 1}{h(G^*) + 1}.$$

And the proof is complete after applying Lemma 3.8 to the above equation.



*Proof of Theorem 3.3:* Let  $E' \in \mathcal{E}$  and  $\mathcal{F}$  be a uniformly random forest of  $G(E')$ . By combining Lemma 3.7 and Lemma 3.2,  $\mathbb{E}[\phi_u(\mathcal{F})] \geq 1 - (\ln n + 2)/(h(G_S) + 1)$ . This inequality, combined with Lemma 3.6, implies that  $\Phi_{u,S}(G) \geq 1 - \frac{2}{h(G_S)}$ .

We conclude this section with a conjecture on RF-connectivity of a vertex in the Erdős-Rényi graph. Let  $G(n, p)$  denote a random graph over  $n$  nodes where every edge is present with probability  $p$ , and let  $V$  be the vertex set of  $G$ . We denote  $\Phi_{u,V}(G)$  by simply  $\Phi_u(G)$  below.

**CONJECTURE 3.1.** *Consider a random graph  $G(n, p)$  where  $p > \frac{2 \ln n}{n}$ . There exists a constant  $c$  and a function  $\varepsilon(n)$ , such that  $\Pr(\Phi_u(G) \geq 1 - \frac{c}{np}, \forall u \text{ in } G) \geq 1 - \varepsilon(n)$ , where  $\varepsilon(n)$  goes to 0, as  $n$  goes to infinity.*

We believe that the key to resolving this conjecture is to better understand the structure of a uniformly random spanning tree of  $G(n, p)$ . We know that when  $p > 2 \ln n/n$ , the diameter of a uniformly random spanning tree of  $G$  is  $O(\sqrt{n} \log n)$  [8, 18], with probability tending to 1 as  $n$  becomes large. However, not much improvement can be achieved by using this information alone in the proof of Lemma 3.7. In particular, any further improvements seem to require understanding the distribution of degree 1 and 2 nodes in a random spanning tree as well as the expected sizes of subtrees when a random edge is deleted. We would like to note that these questions are of independent interest [26].

#### 4 Negative Correlation and Monotonicity of Connectivity

We will now define negative correlation and monotonicity of connectivity property of graphs and show (proofs deferred to the appendix) that negative correlation implies that connectivity is monotone. We begin by defining the negative correlation property. Let  $\mathcal{F}$  be a forest chosen uniformly at random from  $\mathcal{F}$ . The probability that  $X \subseteq E$  is present in  $\mathcal{F}$  is given by  $P_X(G)$ . If  $\mathcal{F}_X$  be the largest subset of  $\mathcal{F}$  with the property that  $X$  is included in every forest of  $\mathcal{F}_X$ , then,

$$P_X(G) = \frac{|\mathcal{F}_X|}{|\mathcal{F}|}.$$

**Negative correlation property:** Given a graph  $G = (V, E)$ , we say that the edges  $(u, v), (u', v') \in E$  are *negatively correlated* if

$$P_{\{(u,v),(u',v')\}}(G) \leq P_{(u,v)}(G) \cdot P_{(u',v')}(G).$$

or,

$$(4.4) \quad P_{(u',v')}(G \mid (u, v)) \leq P_{(u',v')}(G).$$

Here  $P_{(u',v')}(G \mid (u, v))$  is the probability of  $(u', v')$  being present given that the forest is chosen uniformly at random from all forests that have the edge  $(u, v)$ . We say that a graph has *negative correlation* property if all pairs of edges are negatively correlated.

**Monotonicity of connectivity property.** Given a graph  $G = (V, E)$ , let  $(u, v) \notin E$ . Let  $G' = (V, E')$  be such that  $E' \leftarrow E \cup \{(u, v)\}$ . For any pair of vertices  $u, v \in V$ , we say that their RF-Connectivity is monotonically non-decreasing if  $\forall (u', v') \notin E$  and  $E' = E \cup \{(u', v')\}$  with  $G' = (V, E')$ ,  $\Phi_{u,v}(G') \geq \Phi_{u,v}(G)$ . We say that a graph has the monotonicity property if every pair  $(u, v) \in V \times V$  has monotonically non-decreasing RF-Connectivity.

**THEOREM 4.1.** *The negative correlation property holds for all graphs if and only if connectivity is monotone.*

Let  $\mathcal{F}(u, v)$  denote the largest subset of  $\mathcal{F}$  with the property that  $u$  and  $v$  belong to the same component of every forest of  $\mathcal{F}(u, v)$ , then

$$\Phi_{u,v}(G) = \frac{|\mathcal{F}(u, v)|}{|\mathcal{F}|}.$$

The following lemma proves some useful relations:

**LEMMA 4.1.** *Let  $G = (V, E)$  be any graph and let  $G' = (V, E')$  be a graph with  $E' \leftarrow E \cup \{(u, v)\}$  where  $(u, v) \notin E$ . Then,*

$$P_{(u,v)}(G') = \frac{1 - \Phi_{u,v}(G)}{2 - \Phi_{u,v}(G)}$$

$$\Phi_{u,v}(G') = \frac{1}{2 - \Phi_{u,v}(G)}.$$

*Proof.* We use  $\mathcal{F}'$  to denote the set of forests of  $G'$ . Recall that  $\Phi_{u,v}(G)$  is the fraction of forests of  $G$  that have  $u$  and  $v$  in the same component. Therefore,  $1 - \Phi_{u,v}(G)$  is the fraction of forests that do not have  $u$  and  $v$  in the same component. For every forest which has  $u$  and  $v$  in different component, we construct a new forest by adding the edge  $(u, v)$  to it. It is easy to see that the resulting set of new forests is precisely  $\mathcal{F}'_{(u,v)}$ . Therefore,  $\mathcal{F}'$  can be simply obtained taking the union of  $\mathcal{F}$  with  $\mathcal{F}'_{(u,v)}$ . The new probability of connectivity,  $\Phi_{u,v}(G')$  is given by

$$(4.5) \quad \Phi_{u,v}(G') = \frac{|\mathcal{F}(u, v)| + |\mathcal{F}'_{(u,v)}|}{|\mathcal{F}| + |\mathcal{F}'_{(u,v)}|} = \frac{1}{2 - \Phi_{u,v}(G)}.$$

The above equality is obtained by dividing the numerator and denominator by  $|\mathcal{F}|$  and the fact that  $\frac{|\mathcal{F}'_{(u,v)}|}{|\mathcal{F}|} =$

$1 - \Phi_{u,v}(G)$ . Similarly, we also get the probability that the edge  $(u, v)$  is present in a random forest of  $G'$  as

$$(4.6) \quad P_{(u,v)}(G') = \frac{|\mathcal{F}'_{(u,v)}|}{|\mathcal{F}| + |\mathcal{F}'_{(u,v)}|} = \frac{1 - \Phi_{u,v}(G)}{2 - \Phi_{u,v}(G)}$$

*Proof of Theorem 4.1:*

First, assuming negative correlation property, we show that the all pairs of vertices have non-decreasing monotonicity.

We begin by showing that all pairs that are non-adjacent in  $G$  i.e., all pairs  $(u, v) \notin E$ , have the monotonicity property. Next, we use an easy extension to prove monotonicity for all adjacent pairs as well.

We will now show that monotonicity holds for a pair of non-adjacent vertices. Let  $u, v$  be any pair of non-adjacent vertices, and let  $(u', v')$  be the edge that is added to  $G$  to obtain  $G'$ . We augment the graph  $G'$  by adding the edge  $(u', v')$  to it; let the resulting graph be  $G''$ . There are two cases:

1. First, the pair  $(u', v')$  is the same as the pair  $(u, v)$ . In this case, we need to show that adding the edge  $(u, v)$  to  $G$  increases the connectivity of  $u$  and  $v$ . From equation (4.5), for all  $0 \leq \Phi_{u,v}(G) \leq 1$ , we get

$$\Phi_{u,v}(G') \geq \Phi_{u,v}(G).$$

Next, we prove the claim for all non-adjacent pairs  $(u, v)$  where  $(u, v)$  and  $(u', v')$  are distinct. From the negative correlation in  $G''$ , we have

$$P_{(u,v)}(G'' \mid (u', v')) \leq P_{(u,v)}(G'').$$

We will now show that  $\Phi_{u,v}(G) \leq \Phi_{u,v}(G')$ , i.e., adding the edge  $(u', v')$  to  $G$  will only increase the RF-Connectivity of  $u$  and  $v$ .

From (4.6), we have

$$(4.7) \quad P_{(u,v)}(G'') = \frac{1 - \Phi_{u,v}(G')}{2 - \Phi_{u,v}(G')}.$$

Next, suppose we restrict ourselves to all the forests of  $G''$  that contain the edge  $(u', v')$ <sup>4</sup>. An analysis similar to the proof of equation (4.6), we get

$$(4.8) \quad P_{(u,v)}(G'' \mid (u', v')) = \frac{1 - \Phi_{u,v}(G' \mid (u', v'))}{2 - \Phi_{u,v}(G' \mid (u', v'))}.$$

<sup>4</sup>We extend all the definitions to this set. We use the notation  $G|(u', v')$  to indicate that the definition is restricted to the set of forests containing the edge  $(u', v')$

Applying (4.7) and (4.8) to (4.4) we get,

$$\begin{aligned} \Phi_{u,v}(G') &\leq \Phi_{u,v}(G' \mid (u', v')) \\ \frac{|\mathcal{F}(u, v)| + |\mathcal{F}'_{(u',v')}(u, v)|}{|\mathcal{F}| + |\mathcal{F}'_{(u',v')}|} &\leq \frac{|\mathcal{F}'_{(u',v')}(u, v)|}{|\mathcal{F}'_{(u',v')}|} \end{aligned}$$

Here,  $\mathcal{F}'_{(u',v')}(u, v)$  is the set of forests of  $G'$  that contain the edge  $(u', v')$  and have  $u$  and  $v$  in the same component. The previous equation can be re-written as,

$$\begin{aligned} \frac{|\mathcal{F}(u, v)|}{|\mathcal{F}|} &\leq \frac{|\mathcal{F}(u, v)| + |\mathcal{F}'_{(u',v')}(u, v)|}{|\mathcal{F}| + |\mathcal{F}'_{(u',v')}|} \\ \implies \Phi_{u,v}(G) &\leq \Phi_{u,v}(G') \end{aligned}$$

Now we describe the case where  $u$  and  $v$  are adjacent in graph  $G$ .

2. For the case where  $(u, v)$  is an adjacent pair. Let  $\hat{G}$  be the graph obtained by removing the edge  $(u, v)$ . Let  $\hat{G}'$  be the graph after adding edge  $(u', v')$  to  $\hat{G}$ . Since  $u$  and  $v$  are non-adjacent, from the proof of case(ii), we know that  $\Phi_{u,v}(\hat{G}) \leq \Phi_{u,v}(\hat{G}')$ . Next, we add the edge  $(u, v)$  to both  $\hat{G}$  and  $\hat{G}'$  to obtain  $G$  and  $G'$ . From equation (4.5), we get

$$\begin{aligned} \Phi_{u,v}(G) &= \frac{1}{2 - \Phi_{u,v}(\hat{G})} \\ \Phi_{u,v}(G') &= \frac{1}{2 - \Phi_{u,v}(\hat{G}')} \end{aligned}$$

From the above equations and the fact that  $\Phi_{u,v}(\hat{G}) \leq \Phi_{u,v}(\hat{G}')$ , we get  $\Phi_{u,v}(G) \leq \Phi_{u,v}(G')$ .

This concludes the proof in one direction, i.e., given negative correlation holds, we show that connectivity is non-decreasing. Now, we prove the other direction that if connectivity is non-decreasing then the negative correlation property holds.

We prove this for any two edges  $(u, v)$  and  $(u', v')$  in  $G''$ . Given  $(u, v)$  and  $(u', v')$ , let  $G, G'$  and  $G''$  as defined before. From the monotonicity property,  $\Phi_{u,v}(G) \leq \Phi_{u,v}(G')$ , we have

$$\Phi_{u,v}(G') \leq \Phi_{u,v}(G' \mid (u', v'))$$

Note that  $G''$  is obtained by adding the edge  $(u, v)$  to  $G'$ . From (4.7) and (4.8), we get

$$(4.9) \quad P_{(u,v)}(G'') \geq P_{(u,v)}(G'' \mid (u', v')).$$

But,  $P_{(u,v)}(G'' \mid (u', v')) = \frac{P_{(u,v),(u',v')}(G'')}{P_{(u',v')}(G'')}$ . Plugging this to (4.9), we get

$$P_{(u',v')}(G'')P_{(u,v)}(G'') \geq P_{(u,v),(u',v')}(G'')$$

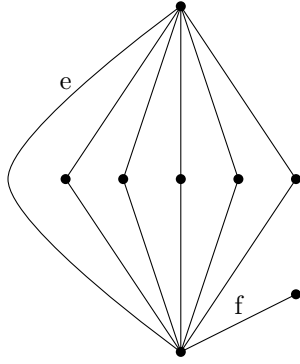


Figure 1: Consider the set of forests with six edges. It can be verified that  $\Pr(e,f) = \frac{5}{24} > \Pr(e)\Pr(f) = \frac{7}{24} \times \frac{17}{24}$ .

implying negative correlation of  $(u, v)$  and  $(u', v')$ .

**Remark:** We know that the negative correlation property does not extend to the truncation of the set of forests by a fixed number of edges (a counter example is already implicitly mentioned in [14]. We draw it in Figure 1 for the completeness of our paper). We also know that the property holds when  $G$  is a complete graph [29] or a series-parallel graph [31]. And the conjecture has been verified when  $G$  has eight or fewer vertices [17]. However, not much is known beyond that.

## Acknowledgements

We are grateful to Chandra Chekuri and Pranav Dandekar for many useful discussions in the early stages of this work.

## References

- [1] D. Aldous. The random walk construction of uniform spanning trees and uniform labelled trees. *SIAM J. Discrete Math.*, 3(4):450465, 1990.
- [2] J. Annan. A randomised approximation algorithm for counting the number of forests in dense graphs. *Combinatorics, Probability and Computing*, 3(03):273–283, 1994.
- [3] C. M. Bishop. *Pattern recognition and machine learning*. springer New York, 2006.
- [4] B. Bollobás. *Random graphs*. Springer, 1998.
- [5] A. Bonato. A survey of properties and models of online social networks. In *Proc. of the 5th International Conference on Mathematical and Computational Models, ICMCM*, 2009.
- [6] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [7] A. Broder. Generating random spanning trees. *Proceedings of the 30th IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 442–447, 1989.

- [8] F. Chung, P. Horn, and L. Lu. Diameter of random spanning trees in a given graph. *Journal of Graph Theory*, 69(3):223–240, 2012.
- [9] H. Dai. Perfect sampling methods for random forests. *Advances in Applied Probability*, pages 897–917, 2008.
- [10] P. Dandekar. *Trust and Mistrust in a Networked Society*. PhD thesis, Stanford University, 2013. Available at [http://web.stanford.edu/~ppd/papers/ppd\\_dissertation.pdf](http://web.stanford.edu/~ppd/papers/ppd_dissertation.pdf).
- [11] P. Dandekar, A. Goel, R. Govindan, and I. Post. Liquidity in credit networks: A little trust goes a long way. In *Proceedings of the 12th ACM conference on Electronic commerce*, pages 147–156. ACM, 2011.
- [12] P. Dandekar, A. Goel, M. Wellman, and B. Wiedendek. Strategic formation of credit networks. *Proceedings of the 21st International World Wide Web conference (www2012)*, 2012.
- [13] D. B. DeFigueiredo and E. T. Barr. Trustdavis: A non-exploitable online reputation system. In *CEC '05: Proceedings of the Seventh IEEE International Conference on E-Commerce Technology*, pages 274–283, Washington, DC, USA, 2005.
- [14] T. Feder and M. Mihail. Balanced matroids. In *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*, pages 26–38. ACM, 1992.
- [15] A. Ghosh, M. Mahdian, D. M. Reeves, D. M. Pennock, and R. Fugger. Mechanism design on trust networks. In *WINE '07: Proceedings of the 3rd international workshop on Internet and Network Economics*, pages 257–268, 2007.
- [16] E. Gioan. Enumerating degree sequences in digraphs and a cycle–cocycle reversing system. *European Journal of Combinatorics*, 28(4):1351–1366, 2007.
- [17] G. R. Grimmett and S. N. Winkler. Negative association in uniform forests and connected graphs. *Random Struct. Algorithms*, 24(4):444–460, 2004.
- [18] C. Hoffman, M. Kahle, and E. Paquette. Spectral gaps of random graphs and applications to random topology. *arXiv preprint arXiv:1201.0425*, 2012.
- [19] J. Kahn. A normal law for matchings. *Combinatorica*, 20(3):339–391, 2000.
- [20] D. J. Kleitman and K. J. Winston. Forests and score vectors. *Combinatorica*, 1(1):49–54, 1981.
- [21] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th international conference on World Wide Web*, pages 695–704. ACM, 2008.
- [22] T. Luczak. Phase transition phenomena in random discrete structures. *Discrete Mathematics*, 136(1-3):225242, 1994.
- [23] R. Lyons and Y. Peres. *Probability on Trees and Networks*. In preparation. Current version available at <http://mypage.iu.edu/~rdlyons>, 2014.
- [24] W. Myrvold. Counting k-component forests of a graph. *Networks*, 22:647652, 1992.
- [25] R. Pemantle. Towards a theory of negative independence. *Journal of Mathematical Physics*, 41:1371–1390,

2000.

- [26] R Pemantle. Uniform random spanning trees. *arXiv preprint math/0404099*, 2004.
- [27] M. Pinsky. On the complexity of a concentrator. In *7th International Teletraffic Conference*, volume 4, pages 1–318, 1973.
- [28] R. Stanley. Decompositions of rational convex polytopes. *Ann. Discrete Math.* v6, pages 333–342, 1980.
- [29] D. Stark. The edge correlation of random forests. *Annals of Combinatorics*, 15(3):529–539, 2011.
- [30] W. T. Tutte. A problem on spanning trees. *Quart. J. Math. Oxford*, 25:253255, 1974.
- [31] D. Wagner. Negatively correlated random variables and mason’s conjecture for independent sets in matroids. *Annals of Combinatorics*, 12(2):211–239, 2008.
- [32] D. Welsh. The tutte polynomial. *Random Structures & Algorithms*, 15(3-4):210228, 1999.

## A Omitted Details from Section 2.2

Let  $s, s' \in \sigma$  be two states in a credit network  $\mathcal{G} = (\mathcal{V}, \mathcal{E}; c(\cdot))$ . Let  $d_u(s)$  denote the indegree of a node  $u$  in state  $s$  and  $\vec{d}(s)$  denote the indegree sequence associated with  $s$ . Two states  $s$  and  $s'$  are said to be equivalent if and only if they correspond to the same indegree sequence.

Dandekar et al. (Lemma 2.2 in [10]) characterizes that two states  $s$  and  $s'$  are equivalent if and only if the network can be transformed from state  $s$  to state  $s'$  by routing transactions along feasible cycles. This observation implies that the set of transactions that goes through in  $s$  is equal to the set of transactions that goes through in  $s'$ . It also leads naturally to a *path-independence property* (Theorem 2.4 in [10]): a successful payment routing from  $u$  to  $v$  along any path leads to the same equivalence class in the Markov chain. These two properties ensure that the repeated transaction model reduces to a Markov chain over the set of indegree sequences  $\mathcal{S}$ . Under a symmetric transaction regime, the stationary distribution of this Markov chain is uniform over  $\mathcal{S}$  (Corollary 2.10 in [10]). Therefore, if  $\mathcal{S}_{u,v}$  is the set of indegree sequences in which transactions from  $u$  to  $v$  go through, then

$$\Psi_{u,v}(\mathcal{G}) = |\mathcal{S}_{u,v}|/|\mathcal{S}|.$$

It is well-known that the number of feasible indegree sequences of  $\mathcal{G}$  is equal to the number of forests of  $G$  [20, 16, 28]. Moreover, the proofs from [20] implicitly implies Proposition 2.1. We present a brief proof here for the completeness of this section.

*Proof of Proposition 2.1:* Let  $\mathcal{F}(u, v)$  denote the set of forests in which  $u$  and  $v$  are connected. Slightly abusing the notation, for any edge  $e \in E$ , we let  $\mathcal{F}_e$  denote the set of forests that contain the edge  $e$ , and let  $\mathcal{F}_{\bar{e}}$  denote the set of forests that do not contain the edge  $e$ . We will prove that  $|\mathcal{F}(u, v)| = |\mathcal{S}_{u,v}|$ . Since  $|\mathcal{F}| = |\mathcal{S}|$ , therefore  $\Phi_{u,v}(G) = \Psi_{u,v}(\mathcal{G})$ . Consider two cases.

- If there exists a labeled edge  $e$  between  $u$  and  $v$  in  $G$ : Let  $\mathcal{S}' = \{\vec{d}(s) \mid \forall s \in \sigma \text{ s.t. } s(e) = \langle v, u \rangle\}$ , where  $s(e) = \langle v, u \rangle$  means that  $e$  is directed from  $v$  to  $u$  in state  $s$ . It is not hard to see that the amount of indegree sequences in  $\mathcal{S}'$  is equal to the amount of forests in  $G \setminus \{e\}$ . Next, for any state  $s'$  such that  $\vec{d}(s') \in \mathcal{S} \setminus \mathcal{S}'$ , by definition  $e$  is oriented from  $u$  to  $v$ . And by Lemma 2.2 in [10], there is no path from  $v$  to  $u$  in  $s'$ . In other words, a transaction  $\langle v, u \rangle$  cannot go through in  $s'$ . Hence  $\mathcal{S}_{u,v}$  is equal to the set  $\mathcal{S}'$ . Finally, note the simple fact  $|\mathcal{F}(u, v)| = |\mathcal{F}| - |\mathcal{F}_e| = |\mathcal{F}_{\bar{e}}|$ , therefore  $|\mathcal{F}(u, v)| = |\mathcal{S}_{u,v}|$ .
- If there does not exist any labeled edge between  $u$  and  $v$  in  $G$ : Think of adding an edge  $e$  between the two vertices in  $G$  and a unit of credit limit between  $u$  and  $v$  in  $G$ . Let  $G'$  denote the new graph and  $\mathcal{G}'$  denote the new network. Let  $\mathcal{S}''$  denote the set of indegree sequences in  $\mathcal{G}'$  such that the transaction  $\langle u, v \rangle$  does not go through. We would like to show that  $|\mathcal{S}_{u,v}| = |\mathcal{S}| - |\mathcal{S}''|$  and  $|\mathcal{S}''| = |\mathcal{F}_e(G')|$ . Combined with  $|\mathcal{F}(u, v)| = |\mathcal{F}| - |\mathcal{F}_e(G')|$ , we would get  $|\mathcal{F}(u, v)| = |\mathcal{S}_{u,v}|$ . The arguments is quite similar to the first case, so we omit the details here.