First passage percolation on locally treelike networks. I. Dense random graphs

Shankar Bhamidi[a]

Department of Mathematics, University of British Columbia, Room 121,
1984 Mathematics Road Vancouver, British Columbia V6T 1Z2, Canada

We study various properties of least cost paths under independent and identically distributed (iid) disorder for the complete graph and dense Erdős–Rényi random graphs in the connected phase, with iid exponential and uniform weights on edges. Using a simple heuristic, we compute explicitly limiting distributions for (properly recentered) lengths of shortest paths between typical nodes as well as multiple source destination pairs; we also derive asymptotics for the number of edges on the shortest path, namely, the *hopcount*, and find that the addition of edge weights converts these graphs from *ultrasmall* world networks to *small* world networks. Finally we study the Vickrey–Clarke–Groves measure of overpayment for the complete graph with exponential edge weights and show that the complete graph is far from monopolistic for large *n*. © *2008 American Institute of Physics.*
[DOI: 10.1063/1.3039876]

I. INTRODUCTION

The past few years have seen an explosion in the amount of data on *real-world* networks including spatial networks such as rail and road networks and data networks such as the Internet. Data transmission networks such as the Internet, which try to route flow between different parts of the network, have two components to them, the graphical structure representing the actual topology of the network and edge costs representing various notions of cost such as congestion or actual economic costs in routing flow across edges. At the very basic level the modeling of these networks consists of two different aspects:

(a) finding mathematical models which explain various crucial statistics of these networks such as the degree distribution and the clustering coefficient and

(b) models for the edge costs.

The cost of a path in the network is the sum of the cost of the edges on the path. Typically nodes try to communicate through the least cost paths.

In this paper we envisage a situation where, while the edge costs represent actual economic costs in transmitting flow across edges in the network, the actual time it takes for messages to get from one node to another is proportional to the number of edges on this least cost path, namely, the *hopcount*. Thus it is of interest to understand both the actual economic cost of the minimum cost path between nodes in the network as well as the number of edges on this path.

We shall consider some math models of random graphs and analyze the simplest setup where the costs are independent and identically distributed. We shall find asymptotics for both the cost of paths between two typical nodes in the network as well as the number of edges on this path. We shall also find the joint distribution of recentered costs between multiple sources and destinations. Although in all of these graphs, the number of edges in the **graph distance** is much smaller than log *n* where *n* is the number of nodes in the network (such graphs are usually called "ultrasmall"

---

[a]Electronic mail: shankar@math.ubc.ca.

networks), adding independent and identically distributed (iid) random edge costs drastically changes the geometry of the network, and we see universal behavior emerging in the sense that the hopcount is $\approx \log n$ in these models. More precisely we mean that the hopcount H_n satisfies the asymptotics

$$\frac{H_n}{\log n} \xrightarrow{P} C \tag{1}$$

for some model dependent constant $C > 0$. This is part of a general scheme to study the following (vaguely stated) math program: In a wide class of random graph models, whatever the actual nature of the network topology be, the addition of random edge weights forces the asymptotics given by Eq. (1).

We then turn our attention to a more advanced notion, namely, the Vickrey–Clarke–Groves (VCG) measure of overpayment, and specialize to the case of the complete graph. This notion essentially tries to measure the extent of *monopoly* in the network, namely, assuming that each edge of the network is under the control of an individual (we use the term operator), under some notion of "fair" compensation to the operator (for operating that edge), the question that we would like to quantify is as follows: Are there edges in the network for which the operator in charge of that edge can charge an arbitrarily large amount?

We show that in the complete graph at least with exponential edge costs, this is not possible. See Sec. II where we further elaborate on the economic and game theoretic rationale of the VCG mechanism.

Organization of the paper and brief summary of our results: In Sec. II we describe the notation as well as various mathematical constructs such as point processes which are required to state our main results. In Sec. III we state all the main results of this paper. Section IV describes related literature which is closely associated with our study as well as known results. Section V contains proofs of all the results. We conclude in Sec. VI with a wide ranging discussion on the methods used in this paper, conjectures, and comparisons to other known proof techniques.

II. DEFINITIONS

Here we define the essential mathematical constructs and set up notation.

A. Graph theoretical notation

In all our different models, we shall always deal with an edge weighted connected graph denoted as

$$\mathcal{G}_n = (V_n, E_n, (l(e)_{e \in E_n})),$$

where V_n denotes the node set, E_n denotes the edge set, and $l(e) > 0$ denotes the weight assigned to edge e. We use $\mathcal{G}_n \backslash e$ to denote the graph $\mathcal{G}_n$ where the edge e has been deleted from the graph. We shall use weights, lengths, and costs interchangeably when describing the weights attached to edges. We shall often think of the nodes as labeled by the set $[n] := \{1, 2, \ldots, n\}$.

Least cost (shortest) path and associated statistics: The edge costs $(l(e))_{e \in E_n}$ induce a metric on the graph. For a given pair of nodes $(s, t) \in \mathcal{G}_n$ and a path $\pi(s, t)$ between them, let $\mathbf{L}(\pi(s, t))$ denote the cost of the path, namely, the sum of the edge weights on the path. Among all such paths, let $\boldsymbol{\pi}^1(u, v)$ denote the path with the least cost, and let $\mathbf{L}^1_{s,t}(\mathcal{G}_n)$ denote the cost of this path. In general let $\boldsymbol{\pi}^k(s, t)$ be the kth least cost path from s to t, and let $\mathbf{L}^k_{s,t}$ be the corresponding cost of this path. We shall use *least cost* path and *shortest* path interchangeably.

Hopcount: We shall use $H_n(s, t)$ to denote the number of edges on the least cost path $\boldsymbol{\pi}^1(s, t)$. This is often called the hopcount between s and t.

Throughout, let $(Y_i)_{i \geq 1}$ and $(\hat{Y}_i)_{i \geq 1}$ denote independent and identically distributed exponential random variables with mean 1, with i running over some index set.

Random recursive tree: This is a random growing rooted tree which has the following

growth dynamics: Nodes arrive at discrete times. Start with a single node at time 1 which we shall assign to be the root of the tree. After growing a tree on n nodes, at time $n+1$ a new node is born and chooses one of the preceding n nodes uniformly at random to be its ancestor.

B. Vickrey–Clarke–Groves measure of overpayment

We now come to our first nontrivial math model. Suppose the edge weights $l(e)$ are chosen as an iid environment with common probability distribution $\nu(\cdot)$ on $\mathbb{R}^+$. Our standing assumption throughout, unless otherwise qualified, is that the edge weights are independent and identically distributed as exponential random variables i.e., $l(e) \sim_{\text{iid}} \exp(\mu)$ with some rate μ. The *fair value* or worth of an edge e when transporting a unit flow from source s to destination t, with $s, t \in \mathcal{G}_n$, is defined as

$$v(e,s,t) = (l(e) + \mathbf{L}_{s,t}^1(\mathcal{G}_n \setminus e, s, t) - \mathbf{L}_{s,t}^1(\mathcal{G}_n, s, t)) 1_{\{e \in \boldsymbol{\pi}(s,t)\}}.$$

Note that if e is not in the shortest path from s to t, then the relative worth of the edge e is zero. Conceptually, if $l(e)$ is the cost that the network manager incurs whenever a unit flow passes across the edge e, then $v(e,s,t)$ is the amount *per unit flow* he deserves to charge consumers interested in sending flow from source s to destination t with $l(e)$ measuring his operational cost and $(\mathbf{L}_{s,t}^1(\mathcal{G}_n \setminus e, s, t) - \mathbf{L}_{s,t}^1(\mathcal{G}_n, s, t))$ measuring the excess *profit* or *bonus* for operating the particular edge.

Remark: If the deletion of an edge e causes two pairs of nodes to get disconnected, then the usual convention is to take the distance in the edge deleted graph to be infinity. However, the convention we use is that in this case $\mathbf{L}_{s,t}^1(\mathcal{G}_n \setminus e, s, t) = \mathbf{L}_{s,t}^1(\mathcal{G}_n, s, t)$, the conceptual idea being that this node is very precious so it cannot be deleted. If it was in the hands of a private operator then he could charge as much as he wants. Thus in this setup we assume that the edge is operated by some subsiding authority charging only the operational cost $l(e)$, for example, the government.

Now assuming unit demand rate between every pair of nodes $s, t \in \mathcal{G}_n$, the *total cost* incurred by the network manager in transporting this flow is

$$\text{TC} := \sum_{(s,t) \in \mathcal{G}_n} \mathbf{L}_{s,t}^1(\mathcal{G}_n).$$

The *total price* TP paid by the customers for transporting this flow is

$$\text{TP} := \sum_{(s,t) \in \mathcal{G}_n} v(e,s,t).$$

Thus the total bonus being paid to the network operator is

$$\text{TB} := \text{TP} - \text{TC}.$$

We are interested in measuring the extent of monopoly in this setup, namely, are there edges which are so precious that the above quantity is very large, i.e., deletion of edge e causes a major rerouting of flow between s and t. In such cases the operator of that edge can charge a very large amount for such edges.

To make this idea more precise, define the quantity $\rho(\mathcal{G}_n)$, the expected VCG measure of overpayment, as

$$\rho(\mathcal{G}_n) = \frac{\mathbb{E}_\nu(\text{TB})}{\mathbb{E}_\nu(\text{TC})}, \tag{2}$$

where the expectation is taken both over the random graph model and the subsequent edge disorder.

Remark: At a probabilistic level it might make more sense to look at just the fraction TB/TC and it is quite conceivable that it should be possible to show limit theorems for this particular

construct at least in a simple setting such as the complete graph with exponential mean 1 edge weights. In this setting it should be fairly easy to show that

$$\frac{\text{TB}}{n \log n} \to 1$$

in probability as $n \to \infty$. However, dealing with the numerator seems far more challenging. Thus we restrict ourselves to the statistic given by Eq. (2).

   Economic rationale for VCG: Note that the above in some sense quantifies the *value* of different edges in the network. Edges which are in the least cost paths of many source destination pairs and whose removal from the network would cause many source destination pairs to use much longer and more circuitous routes are paid more compared to other edges.

   Game theoretic rationale for VCG: Many of the fundamental algorithms in computer science are geared up to find short paths in networks. In the design of such algorithms, there is an implicit assumption that the nodes will act as instructed, unless they are faulty or malicious. However, with the phenomenal growth of the Internet, the above is not necessarily a valid assumption. Various parts of the Internet are owned by different Internet service providers, each interested in maximizing their own net economic gain or profit.

   Consider the problem that is the main theme of this paper, namely, shortest paths in the network. Suppose for each edge e that the maintenance cost (that we model as iid random variables) is some private quantity which the operator does not want to *a priori* share with the rest of the network. For ease of exposition we shall stick to the single source destination setup. Fix two nodes x and y. Our goal is to route flow from x to y through the *true* least cost path and pay the operators of edges some fair compensation. Given some claims, say $(l'(e))_{e \in E_n}$, that operators make about the costs on their edges our algorithm would be the following:

(a)   Compute the least cost path in the network using the cost environment proposed by the operators.
(b)   Give the operators on edges **on the shortest path** some compensation $p(e, \{l(e')\}_{e' \in E_n})$. Note that an edge not on the shortest path does not get any compensation. The final true utility of the operator controlling edge e is

$$v(e,p) = [p(e, \{l(e')\}_{e' \in E}) - l(e)] 1\{e \text{ is paid}\}.$$

   Note that in the above we have subtracted their true operating cost $l(e)$. Also note the notation $1\{\cdot\}$ for the indicator operator of a set.

   We would like to design a payment mechanism $p(\cdot)$ which would force the operators to reveal their true operating costs. We assume that edges always try to act in a manner so as to maximize their utility. We also assume that edges do not propose edge costs less than their true edge costs. Let $\mathbf{L}_{x,y}(\mathcal{G}_n, (l'(e)))$ denote the least cost path between the two nodes x and y when the proposed environment is $\{l'(e)\}$.

   The VCG mechanism which we once again state is as follows: Given the proposed edge costs $l'(e)$, give each e (which is present on the least cost path from x to y) an amount

$$p(e, \{l(e')\}_{e' \in E}) = l'(e) + \mathbf{L}_{x,y}(\mathcal{G}_n/e, (l'(e)) - \mathbf{L}_{x,y}(\mathcal{G}_n, (l'(e)) \quad .$$

Without going into too much detail, the bottom line is that, under a general game theoretic model, the above mechanism forces all the operators to reveal their true operating costs. Intuitively an operator would not propose a very high edge cost because then it would probably force his edge not to lie on the shortest path between x and y, thus resulting in a zero utility. See Ref. 9 for more details.

C. Random graph models

Here we briefly describe the random graphs we analyze.

Complete graph: This is just the graph nodes on n nodes where each node is attached to every other node via a direct edge.

Erdős–Rényi random graph: Consider the Erdős–Rényi random graph $\mathcal{G}_{n,p_n}$ where we first start with the complete graph and retain each edge with probability p_n and removed with probability $1 - p_n$. In this paper we shall always assume the connected regime, namely, we shall assume that the sequence of connection probabilities p_n is such that $\exists\, a > 1$ such that

$$\liminf_{n \to \infty} \frac{n}{\log n} p_n = a > 1. \tag{3}$$

Although known, we shall give a simple proof that in the above regime

$$\mathbb{P}(\mathcal{G}_{n,p_n} \text{ is connected}) \to 1 \tag{4}$$

as $n \to \infty$.

D. Point processes

Cox point processes on $\mathbb{R}$: Let Z be a positive random variable. Say that a point process $\Xi_Z(\cdot)$ on $\mathbb{R}$ is a Cox point process with the intensity function Ze^x if, conditional on Z, $\Xi_Z(\cdot)$ is a Poisson process on $\mathbb{R}$ with intensity function Ze^x.

E. Yule process

Definition 1: *The standard Yule process is a continuous time Markov process which is defined as follows: We start with one individual at time* 0*. This individual lives forever and reproduces at times of a unit rate Poisson process. Each offspring also lives forever and reproduces at times of independent Poisson processes.*

See Sec. V A where we give some important properties of the Yule process and describe how it ties up to our flow processes on the complete graph.

III. RESULTS

We shall now state our results.

A. Complete graph

Theorem 2 (two point distances): *Consider the complete graph $\mathcal{G}_n$ with each edge independent and identically distributed as exponential random variables with mean 1. Fix nodes 1 and 2. Recall that $\mathbf{L}_{1,2}^1 < \mathbf{L}_{1,2}^2, \ldots$ denote the (actual) costs of the minimal cost paths from node 1 to node 2, arranged in increasing order, and H_n denotes the number of edges on the least cost path, namely, the hopcount. Then*

(a)

$$n\mathbf{L}_{1,2}^1 - \log n \xrightarrow{d} \Xi_{W_1 \cdot W_2}^1 \tag{5}$$

as $n \to \infty$. Here $\Xi_{W_1 \cdot W_2}^1$ denotes the first point in a Cox point process as defined in Sec. II *and W_1, W_2 are independent identically distributed as $\exp(1)$ random variables. In particular, since $\Xi_{12}^1 = \xi_1 + \eta_{12}$, where ξ_1 has a double exponential distribution and η_{12} has a logistic distribution independent of ξ_1 , we have*

$$n\mathbf{L}_{1,2}^1 - \log n \xrightarrow{d} \xi_1 + \eta_{12}. \tag{6}$$

The same fact holds if all the edges are iid $U[0,1]$ *random variables. In particular, for the complete graph with uniform edge weights, we have*

$$\limsup_{n\to\infty} \frac{\max_{e\in\boldsymbol{\pi}(1,2)} l(e)}{\log n/n} \le 1,\tag{7}$$

with probability converging to 1 *as* $n\to\infty$. *Here* $l(e)$ *is the cost of the edge* e *and as before* $\boldsymbol{\pi}(1,2)$ *denotes the least cost path between nodes* 1 *and* 2.

(b)   *For the hopcount we have*

$$\frac{H_n}{\log n} \to_P 1\tag{8}$$

as $n\to\infty$.

(c)   *Fix* $a>1$. *Let* $\mathcal{T}_1$ *be the shortest path tree from* 1 *to all other nodes. Then*

$$\mathbb{P}\left(\exists\ \text{edge}\ \mathbf{e}\in\mathcal{T}_1, l(\mathbf{e}) > \frac{a\log n}{n}\right) \to 0\tag{9}$$

as $n\to\infty$.

Remark: Equation (6) has been derived before although proved via conceptually completely different methods, see Ref. 6. We reprove this result because it is the simplest example exhibiting the crucial conceptual ideas that also form the basis of the more complicated results below. The characterization given by Eq. (5) is new.

Theorem 3 (single source, multiple destinations): *We have the following results for the joint distribution for multiple source destination shortest paths. As before we are in the setting of the complete graph where all edges have exponential mean* 1 *edge lengths.*

(a)   *Fix a single source node, say,* 1, *and destinations* $2,3,\ldots,k$. *Let* $\mathbf{L}_{1,j}^1$ *denote the least cost path to node* j . *Then*

$$(n\mathbf{L}_{1,2}^1 - \log n, \ldots, n\mathbf{L}_{1,k}^1 - \log n) \overset{d}{\to} (\xi_1 + \eta_{12}, \ldots, \xi_1 + \eta_{1k}),\tag{10}$$

where ξ_1 *has the double exponential distribution*

$$\mathbb{P}(\xi \le x) = \exp(-e^{-x}), \quad -\infty < x < \infty,$$

the η_{1j} *have logistic distribution*

$$\mathbb{P}(\eta \le x) = \frac{e^x}{1+e^x}, \quad -\infty < x < \infty,$$

and all the random variables (RV's) in the limits are independent.

(b)   *For general source destination pairs* i *and* j , *we have the following result for the array* $\mathbf{L}^1(i,j)$:

$$(n\mathbf{L}^1(i,j) - \log n, 1 \le i < j \le k) \overset{d}{\to} (D(i,j), 1 \le i < j \le k),\tag{11}$$

where the joint distribution of the limit is

$$(D(i,j), 1 \le i < j \le k) \overset{d}{=} (\xi_i + \xi_j - \xi_{ij}, 1 \le i < j \le k),$$

where all the limit RV's have the double exponential distribution.

Theorem 4: *Let* $\mathcal{G}_n$ *denote the complete graph with iid exponential edge disorder. Let* $\rho(\mathcal{G}_n)$ *denote the VCG measure of overpayment. Then,*

$$\lim_{n \to \infty} \rho(\mathcal{G}_n) \to 0.$$

B. Dense Erdős–Rényi random graph

We now state our results for the dense Erdős–Rényi random graph setup. Surprisingly we can show near optimal results quite easily for the setup where the edge distributions are uniform; however, for the hopcount under the exponential distribution, we have to be a bit more careful.

First assume that the sequence p_n satisfies the following condition:

$$\liminf_{n \to \infty} p_n \frac{n}{\log n} = a, \tag{12}$$

where $1 < a \le \infty$.

Theorem 5: *Consider the Erdős–Rényi random graph $\mathcal{G}_{n,p_n} = \mathrm{ER}(n, p_n)$ with iid uniform edge weights $U[0,1]$. Assume that condition Eq. (12) is met. Then*

(a) $\mathbb{P}(\mathcal{G}_{n,p_n} \text{ is connected}) \to 1$ *as* $n \to \infty$.

(b)   *Fix two node,s say,* 1 *and* 2. *Then*

$$n p_n \mathbf{L}_{1,2} - \log n \xrightarrow{d} \Xi^1_{W_1 \cdot W_2},$$

   where W_1 and W_2 are independent identically distributed as exponentials with mean 1.

(c)   *For the hopcount H_n between the two nodes we have*

$$\frac{H_n}{\log n} \to_P 1$$

   as $n \to \infty$.

Remark: Part (a) of the above theorem is very well known, often called the *connectivity threshold* of the random graph. We state it as a separate theorem here as it follows almost immediately from Theorem 2 part (c) and plays a key role in the proof of the remaining statements of the theorem. An immediate corollary of this result is the following observation.

Corollary 6: *Consider the Erdős–Rényi random graph $\mathcal{G}_n(n, a \log n/n)$, with iid $U[0,1]$ edge weights and $a > 1$. Then*

$$\log n(a \mathbf{L}^1_{12} - 1) \xrightarrow{d} \Xi^1_{W_1 \cdot W_2}$$

as $n \to \infty$.

The setting where we have exponential edge weights is slightly more complex. It requires that p_n permits the graph to be *slightly more dense* than the connected regime, namely,

$$\frac{n}{\log^2 n} p_n \to \infty. \tag{13}$$

Theorem 7: *Assume that p_n satisfies the condition of Eq. (13). Assume that all the edges have iid $\exp(1)$ distribution. Then*

(a)   *For the minimum cost path we have*

$$(n p_n \mathbf{L}^1_{12} - \log n) \xrightarrow{d} \Xi_{W_1 \cdot W_2}.$$

(b)   *For the hopcount H_n we have*

$$\frac{H_n}{\log n} \to_P 1$$

as $n \to \infty$.

Remarks:

(1)   Simple branching process-type arguments tell us that in the Erdős–Rényi $(n, a \log n / n)$, the graph distance between two typical nodes is $O(\log n / \log \log n)$. Thus introducing random disorder into the graph changes it from an ultrasmall world network to a small world network. At first sight this might seem surprising but see Theorem 2 part (b) where the typical hopcount for the complete graph increases from 1 to $\log n$ due to the introduction of random disorder.

(2)   We now believe that this theorem can be strengthened to the regime where we have $np_n \to \infty$, which is the regime where two random nodes 1 and 2 are connected with high probability. This shall be proved elsewhere.

IV. RELATED LITERATURE

First passage percolation, especially on the integer lattice, has been extensively studied for the past 50 years, see Ref. 12 and the more recent survey in Ref. 5. The work closest to our setting is Ref. 6 where Eq. (6) is proved and so is Theorem 2 part (b). Similar ideas were also explored in Ref. 15. However, the conceptual ideas in this paper are slightly different and allow us to get the alternative characterization given by Eq. (5) which also helps pave the way for the more advanced asymptotics of Theorem 3.

Regarding results for the dense Erdős–Rényi random graphs, the asymptotics for the lengths of paths, namely, Theorem 5 part (a) and Theorem 7 part (a), are completely new. The hopcount result for when the distribution is $U[0,1]$, namely, Theorem 5 part (b), is new as far as we know. The hopcount result for the Erdős–Rényi random graph when the edge weights are exponentially distributed were proved in Ref. 14 under stronger conditions (namely, $np_n / \log^3 n \to \infty$). In this study using different methods we extend their results to weaker assumptions on p_n.

Expanding neighborhood techniques for random graphs in the discrete setting have been used extensively to explore shortest path structures and other properties of locally treelike graphs. See Refs. 17, 13, and 16 where the authors carried out an extensive and remarkably complete study of the configuration model. In some sense, our main flow expansion techniques are continuous time extensions of their neighborhood expansions. Also see Ref. 1 where more intricate computations are carried out to study "edge flows" on the complete graph. Finally see Ref. 18 where relations between the random assignment problem and shortest path problem on the complete graph are explored.

Regarding the basic statement and formulation of the VCG measure of overpayment in random graphs, we were first made aware of it by the paper in Ref. 7. In the paper via simulation evidence they conjecture that for the complete graph with (uniform) random edge weights, the VCG measure of overpayment $\rho(\mathcal{G}_n) = \Theta(\log^{1.5} n)$. Under slightly different assumptions (namely, each edge having an exponential distribution) we see that this is not the case; rather we have proved that $\rho(\mathcal{G}_n) \to 0$ as $n \to \infty$. A similar result should be true under the assumption of uniform edge weights but is probably more difficult to prove. Thus the complete graph with random (exponential) edge weights is still very fair and nonmonopolistic.

V. PROOFS

A. Proof of Theorem 2

Part of Theorem 2 part (a): Let $\mathcal{G}_n$ denote the complete graph with iid exponential edge weights. By rescaling, we shall assume that all edge weights are exponentially distributed with mean n (namely, rate $1/n$). Then it is enough to prove that

$$\mathbf{L}_{12}^1 - \log n \xrightarrow{d} \Xi_{W \cdot W_2}^1.$$

We start by collecting some standard properties of the random network $\mathcal{G}_n$. For fixed n and $t \geq 0$ define

$$N_n^{(1)}(t) \coloneqq \text{number of nodes with distance } t \text{ from node } 1,$$

where we include node 1 when performing the above count. Let

$$S_{n,k} \coloneqq \min\{t : N_n(t) = k\}, \quad 1 \leq k \leq n-1,$$

so that $S_{n,k+1}$ is the distance from node 1 to the kth nearest neighbor. Then in this scaling, it is easy to see that

$$(S_{n,k+1} - S_{n,k}, 1 \leq k \leq n-1) \text{ are independent exponential}\left(\frac{k(n-k)}{n}\right). \tag{14}$$

Since the distance between the nodes $\mathbf{L}_{12}^1$ is distributed as $S_{n,V}$ where V is uniform on $\{2, 3, \ldots, n\}$, this fact helped Janson conclude in his very influential paper[6] that

$$\mathbf{L}_{12}^1 - \log n \xrightarrow{d} \xi_1 + \eta_{12},$$

where ξ_1 has the double exponential distribution and η_{12} has the logistic distribution.

However, this method does not seem robust enough to deal with other models and hence we use separate tools to prove this fact which gives the alternate characterization:

$$\Xi_{W_1 \cdot W_2}^1 \overset{d}{=} \xi_1 + \eta_{12}.$$

There is a natural mental picture of first passage percolation, in which at time 0 there is water at node 1 only, and the water spreads along edges at speed 1. Then $N_n^{(1)}(t)$ is the number of nodes wetted by time t.

We begin the proof by stating and proving some strong results connecting the first passage percolation process $N_n^{(1)}(t)$ and how it resembles very closely the Yule process as defined in Sec. II E.

Before we start on the main result, we collect some well known properties of the Yule process. First write S_k for the amount of time required for the time of birth of the kth individual in the population, with $S_1 = 0$. Also define $N(t)$ as the number of nodes born by time t in the Yule process. Then by the definition of the process it is easy to verify that

$$(S_{k+1} - S_k, k \geq 1) \text{ are independent exponential}(k) \text{ r.v.} \tag{15}$$

Compare with the process $(S_{n,k})_{1 \leq k \leq n-1}$ of Eq. (14) to see that in some sense, the Yule process is the large n limit of the $N_n(\cdot)$ process. The following result on the asymptotic properties of the Yule process is one of the main tools we require later to understand the rate of growth of the flow processes about nodes.

Proposition 8: *Fix a sequence $\omega_n \to \infty$ such that $\omega_n = o(\log n)$. Let $N(\cdot)$ be the standard Yule process. Then there exists a random variable $W \sim \exp(1)$ such that*

$$\sup_{t \geq \omega_n} \left| \frac{N(t)}{e^t} - W \right| \to_P 0.$$

Proof: The fact that the Yule process converges to a limiting random variable W follows since $e^{-t}N(t)$ is an $\mathbb{L}^2$ bounded martingale. Doob's $\mathbb{L}^2$ maximal inequality gives the convergence in the above proposition. Finally the fact that the limiting random variable has an exponential distribution follows from the fact that the Yule distribution at time t has a geometric distribution with

parameter e^{-t}. Using Laplace transforms and taking limits gives us that the limiting random variable has an exponential distribution. See Ref. 10 for more details. $\square$

Now we shall show that the first passage percolation process $N_n^{(1)}(\cdot)$ also has a similar behavior, at least for times less than $\log n$.

Proposition 9: *Fix a sequence* $\omega_n = o(\log n) \to \infty$. *Then there exist random variables* W_n *with* $W_n \sim \exp(1)$ *such that*

$$\sup_{\omega_n \le t \le \log n - \omega_n} \left| \frac{N_n^{(1)}(t)}{e^t} - W_n \right| \to_P 0$$

as $n \to \infty$.

Proof: Let $\mathcal{F}(\cdot)$ be the filtration generated by the process $N(\cdot)$. Note that by the above discussion, $(N_n(\cdot), \mathcal{F}(\cdot))$ is a Markov counting process with rates given by

$$\mathbb{P}(N_n^{(1)}(t + dt) - N_n^{(1)}(t) = 1 | \mathcal{F}(t)) = \frac{N_n^{(1)}(t)(n - N_n^{(1)}(t))}{n} dt + o(dt) \qquad (16)$$

when $N_n^{(1)}(t) \le n$. We shall define an associated Markov counting process $B_n(t)$ with $B_n(0) := 0$, associated filtration, as $\mathcal{B}(t) = \sigma(B_n(s) : s \le t)$. Define the combined filtration

$$\widetilde{\mathcal{F}}(t) := \mathcal{F}(t) \vee \mathcal{B}(t).$$

Then the process $B_n(t)$ is defined as

$$\mathbb{P}(B_n(t + dt) - B_n(t) = 1 | \widetilde{\mathcal{F}}(t)) = \left(B_n(t) + \frac{[N_n^{(1)}(t)]^2}{n} \right) dt + o(dt). \qquad (17)$$

Define $b_n(t) = [N_n^{(1)}(t)]^2 / n$. Let T_n be the stopping time defined as the time required for the process $N_n^{(1)}(t)$ to grow to size n. Finally define the random counting process $\widetilde{N}_n^{(1)}(t)$ as

$$\widetilde{N}_n^{(1)}(t) = N_n^{(1)}(t) + B_n(t) \quad \text{for } t \le T_n,$$

$$\widetilde{N}_n^{(1)}(t) = \sum_{i=1}^{\widetilde{N}_n^{(1)}(T_n)} Y_i(t - T_n) \quad \text{for } t > T_n, \qquad (18)$$

where $Y_i(\cdot)$ is an independent sequence of Yule processes. Then by the construction of the rates, it is easy to check the following.

Lemma 10: *The process* $\widetilde{N}_n^{(1)}(\cdot)$ *is a standard Yule process and hence there exists a random variable* $W_n \sim \exp(1)$ *such that*

$$\sup_{\omega_n \le t \le \log n - \omega_n} \left| \frac{\widetilde{N}_n^{(1)}(t)}{e^t} - W_n \right| \to_P 0$$

as $n \to \infty$.

Note that we are trying to prove asymptotics for the sequence

$$e^{-t} N_n^{(1)}(t) = e^{-t} \widetilde{N}_n^{(1)}(t) - e^{-t} B_n(t)$$

for $t \le T_n$. Thus to prove Proposition 9 it is enough to prove the following lemma.

Lemma 11:

(a)   *For the stopping times* T_n *we have*

$$\mathbb{P}(T_n > \log n - \omega_n) \to 1$$

as $n \rightarrow \infty$ so that with high probability, for $t \leq \log n - \omega_n$, we have

$$\widetilde{N}_n^{(1)}(t) = N_n^{(1)}(t) + B_n(t).$$

(b)   *The contribution of the sequence $e^{-t}B_n(t)$ is negligible in the sense that*

$$\sup_{t \leq \log n - \omega_n} \frac{B_n(t)}{e^t} \rightarrow_P 0$$

as $n \rightarrow \infty$.

Proof: To prove part (a) note that $N_n^{(1)}(t) \leq \widetilde{N}_n^{(1)}(t)$ and since $\widetilde{N}_n^{(1)}(t)$ is the Yule process we have

$$\mathbb{E}(\widetilde{N}_n^{(1)}(t)) = e^t$$

for any fixed time t. By Markov's inequality, writing $t_n = \log n - \omega_n$, note that

$$\mathbb{P}(T_n \leq \log n - \omega_n) \leq \mathbb{P}(\widetilde{N}_n(t_n) \geq n) \leq \frac{\exp t_n}{n} \rightarrow 0$$

as $n \rightarrow \infty$.

Part (b) requires a little more work. Let the event $\Omega_n = \{B_n(\frac{1}{3}\log n) = 0\}$ and note that the two processes $\widetilde{N}_n^{(1)}(\cdot)$ and $N_n^{(1)}(\cdot)$ coincide at least up to time $\frac{1}{3}\log n$ on the set Ω_n. Note that by the rate equation, (17), we have

$$1 - \mathbb{P}(\Omega_n) \leq \int_0^{(1/3)\log n} \mathbb{E}(b_n(t))dt \leq \frac{1}{n}\int_0^{(1/3)\log n} Ce^{2t}dt \rightarrow 0$$

as $n \rightarrow \infty$, where we have used the fact that for a Yule process, $\mathbb{E}(N^2(t)) \leq Ce^{2t}$ for some constant C. Thus $\mathbb{P}(\Omega_n) \rightarrow 1$ as $n \rightarrow \infty$. Also note that $e^{-t}B_n(t)$ is a submartingale since

$$d(e^{-t}B_n(t)) = e^{-t}dB_n(t) - e^{-t}B_n(t)dt$$

so that by the rate equation (17)

$$\mathbb{E}(e^{-t}dB_n(t)|\widetilde{\mathcal{F}}(t)) = e^{-t}b_n(t)dt \geq 0.$$

Thus by Doob's L^1 maximal inequality for submartingales, it is enough to prove that

$$\mathbb{E}(e^{-t_n}B_n(t_n)\mathbb{1}\{\Omega_n\}) \rightarrow 0$$

as $n \rightarrow \infty$. Define $f_n(t) = \mathbb{E}(e^{-t}B_n(t)\mathbb{1}\{\Omega_n\})$. Note that $f_n(\frac{1}{3}\log n) = 0$. Further by the rate equation (17) we have

$$f_n'(t) \leq f_n(t) + a_n(t),$$

where $a_n(t) = \mathbb{E}(b_n(t)) \leq Ce^{2t}/n$. This, in particular, implies that $(e^{-t}f_n(t))' \leq e^{-t}a_n(t)$. Combining, we have

$$e^{-t_n}f_n(t_n) = \int_{(1/3)\log n}^{t_n} Ce^t/ndt \rightarrow 0$$

as $n \rightarrow \infty$, hence proved.   □

Fix two nodes 1 and 2 in $\mathcal{G}_n$. We shall now see how the above propositions essentially allow us to infer Theorem 2 part (a) and in some sense contain the main ideas in proving many of the other results in this paper. Start two first passage percolation flows as described above but now *simultaneously* from node 1 and node 2. Let $\mathcal{F}_1(t)$ and $\mathcal{F}_2(t)$ denote the clusters explored by node

1 and node 2, respectively, by time t. Further let $(\widetilde{\mathcal{F}}(t))_{t \geq 0}$ denote the filtration which contains all the information about the two flow clusters upto time t. We say that a *collision* has occurred between the two flow clusters at some time t if the flow from node 1 sees (hits) some node previously seen by node 2 or vice versa. We modify the two flow processes slightly in the following sense: Every time a collision occurs, for example, the flow from node 1 reaches some node already visited by node 2, then this branch of the flow from 1 is stopped. Let $N_n^{(i)}(t)$ be the number of nodes wetted *first* by the flow from node i by time t. Let S_{12} denote the first time that a collision has occurred. The crucial observation that ties these flow processes with our shortest path problem is that

$$\mathbf{L}_{12}^1 = 2S_{12}. \tag{19}$$

Thus understanding the time of the first collision is equivalent to understanding the shortest (or least cost) path between nodes 1 and 2. For a fixed time t, let $\mathcal{NC}(t)$ denote the event that no collision has taken place between the two flow clusters until time t, namely,

$$\mathcal{NC}(t) := \{S_{12} > t\}.$$

Fix any time B large but finite, independent of n. The following lemma gives the rate at which collisions happen. We shall use $c_n(t)$ to denote the number of collisions seen by time t. Then by the definition of the two flow processes, since at any given time t, there are $N_n^{(1)}(t) \cdot N_n^{(2)}(t)$ edges between the two flow clusters, each having an exponential $(1/n)$ distribution, we have the following lemma.

Lemma 12: *The process $c_n(t)$ is a Markov counting process adapted to the filtration $\mathcal{F}(t)$ with rate function given by*

$$\mathbb{P}(c_n(t+dt) - c_n(t) = 1 | \widetilde{\mathcal{F}}(t)) = \frac{N_n^{(1)}(t) \cdot N_n^{(2)}(t)}{n} dt + o(dt),$$

with the initialization $c_n(0) = 0$.

For future reference we shall let

$$\lambda_n(t) = \frac{N_n^{(1)}(t) \cdot N_n^{(2)}(t)}{n}.$$

The following proposition essentially states that for times less than $\frac{1}{2}\log n + B$, the two flow processes $\mathcal{F}_1(t)$ and $\mathcal{F}_2(t)$ behave like independent Yule processes, namely, $\mathcal{F}_1(t) \approx W_1 e^t$ and $\mathcal{F}_2(t) \approx W_2 e^t$, where W_1 and W_2 are exponential(1) random variables and **independent** of each other.

Proposition 13: *Fix $B > 0$. Let $N_n^{(i)}(t)$ be the number of nodes seen by the pruned flow process $\mathcal{F}_i(t)$. Fix a sequence $\omega_n = o(\log n)$ with $\omega_n \to \infty$. Then there exist independent rate 1 exponential random variables $W_n^{(1)}, W_n^{(2)}$,*

$$\sup_{\omega_n \leq t \leq (1/2)\log n + B} \left| \frac{N_n^{(i)}}{e^t} - W_n^{(i)} \right| \to_P 0$$

as $n \to \infty$.

Proof: The proof is almost identical to the proof of Proposition 9. As in Proposition 9, coupled with the flow process $N_n^{(i)}(t)$ are independent Yule processes $\widetilde{N}_n^{(i)}(t)$ via the compensator counting process $B_n^{(i)}(t)$. The two processes $B_n^{(i)}$ for the corresponding flow process $\mathcal{F}_i(\cdot)$ are "essentially" independent of the rate equations:

$$\mathbb{P}(B_n^{(i)}(t+dt) - B_n^{(i)}(t) = 1 | \widetilde{\mathcal{F}}(t)) = \left(B_n^{(i)}(t) + c_n(t) + \frac{[N_n^{(i)}]^2}{n} \right) dt + o(dt), \tag{20}$$

where $c_n(t)$ is the number of collisions by time t. Now to show that the contribution of the compensator process $B_n(t)$ is negligible in the time interval $[\omega_n, \log n - \omega_n]$, we proceed with the same argument as Proposition 9, with the minor modification of the rate equation (17) by (20). The only change in the argument is that we now bound the contribution of the term $c_n(\cdot)$ in the above rate equation. It is enough to show that $c_n(\frac{1}{2}\log n + B) = O_P(1)$. For this we have the following lemma.

   Lemma 14: *The collision counting process $c_n(\cdot)$ satisfies the following properties:*

(a)   *Let $\Omega'_n = \{c_n(\frac{1}{3}\log n) = 0\}$. Then $\mathbb{P}(\Omega'_n) \to 1$ as $n \to \infty$.*
(b)   *The sequence of random variables $c_n(\frac{1}{2}\log n + B)$ are $\mathbb{L}^2$ bounded and hence tight, namely, given any $\varepsilon > 0$ there exists $K_\varepsilon > 0$ such that for all n*

$$\mathbb{P}\left(c_n\left(\tfrac{1}{2}\log n + B\right) < K_\varepsilon \right) > 1 - \varepsilon.$$

   Proof: Note that the two flow process $(N_n^{(1)}, N_n^{(2)})$ satisfies the stochastic domination property

$$(N_n^{(1)}(\cdot), N_n^{(2)}(\cdot)) \leq_{\mathrm{st}} (N^1(\cdot), N^2(\cdot)),$$

where $N^i(\cdot)$ are independent Yule processes. Thus assume that the counting process $c_n(\cdot)$ has the conditional intensity rate

$$\widetilde{\lambda}_n^*(t) = \frac{N^1(t) \cdot N^2(t)}{n}. \tag{21}$$

Then it is enough to prove that $c_n(\cdot)$ satisfies the properties of the lemma. Part (a) is trivial since by Eq. (21)

$$1 - \mathbb{P}(\Omega'_n) \leq \int_0^{(1/3)\log n} \mathbb{E}(\widetilde{\lambda}_n^*(t)) dt \leq \frac{1}{n} e^{2t_n} \to 0$$

as $n \to \infty$ where $t_n = \frac{1}{3}\log n$.
   Similarly to prove part (b) note that by the rate equation (21) we have

$$\mathbb{E}\left(c_n^2\left(\frac{1}{2}\log n + B\right) \right) = 2\frac{1}{n^2} \int_0^{(1/2)\log n + B} \int_t^{(1/2)\log n + B} \mathbb{E}(N^{(1)}(s)N^{(1)}(t)N^{(2)}(s)N^{(2)}(t)) ds dt.$$

Using the fact that for the standard Yule process $N(\cdot)$ $\mathbb{E}(N(t)) \leq Ce^{2t}$ (for some constant C and $\forall t$) while for any $s < t$ $N(t) \overset{d}{=} \sum_1^{N(s)} Y_i(t-s)$ where $Y_i(\cdot)$ are independent standard Yule processes gives us the inequality

$$\mathbb{E}\left(c_n^2\left(\tfrac{1}{2}\log n + B\right) \right) \leq C^2 e^{2B}$$

for all n. This completes the proof of the lemma.   □

   Using the same argument as Proposition 9, with the minor modification of the rate equation (17) by (20) and using Lemma 14 to bound the contribution of the additional $c_n(\cdot)$ term now completes the proof of Proposition 13.   □

   Now to finish the proof of Theorem 2 part (a), note that we are interested in the recentered process $2S_{12} - \log n$ which has the hazard rate

$$\widetilde{\lambda}_n(s) = \lambda_n\left(\frac{1}{2}s + \frac{1}{2}\log n\right) = e^s \frac{N_n^{(1)}\left(\frac{1}{2}\log 2 + \frac{1}{2}s\right)}{e^{(1/2)\log 2 + (1/2)s}} \frac{N_n^{(2)}\left(\frac{1}{2}\log 2 + \frac{1}{2}s\right)}{e^{(1/2)\log 2 + (1/2)s}}.$$

Finally we use Proposition 13 to conclude that

$$e^{-s}\widetilde{\lambda}_n(s) \to_P W_1 \cdot W_2$$

uniformly on $2\omega_n - \log n \le s \le B$, where W_1, W_2 are independent exponential rate 1 random variables. This implies that $2S_{12} - \log n \overset{d}{\to} \Xi^1_{W_1 \cdot W_2}$, since $\Xi^1_{W_1 \cdot W_2}$ has hazard rate $e^s W_1 W_2$ on $\infty < s < \infty$.

This finishes the proof of Eq. (5). Using this it is not hard to derive the distribution of Ξ^1_{12} and see that it matches the distribution given in Eq. (6). To see this note that

$$\mathbb{P}(\Xi^1_{W_1 \cdot W_2} > s) = \mathbb{E}(e^{-e^s W_1 \cdot W_2}) = \int_0^\infty \frac{e^{-w}}{1 + we^s}dw.$$

Similarly, since $\xi_1 \overset{d}{=} -\log W_1$ where ξ_1 has the logistic distribution and W_1 has the double exponential distribution, we have for η_{12} a logistic random variable independent of ξ_1,

$$\mathbb{P}(\xi_1 + \eta_{12} > s) = \mathbb{E}\left(\frac{1}{1 + e^{s - \log W_1}}\right) = \int_0^\infty \frac{e^{-w}}{1 + we^s}dw.$$

This proves the equivalence of the two distributions.   □

Universality with regards to edge weight distribution: We shall now prove that the exact properties of the edge weight distribution do not matter; all that is important is the behavior around 0. This was originally stated in Ref. 6. We shall give a complete proof since we require this in our study of the Erdős–Rényi random graph. We should state, however, that the main ideas are from Ref. 6.

Proposition 15: *Consider the complete graph with uniform $U[0,1]$ edge weights. Let $\mathbf{L}^{\mathrm{unif}}_{12}$ denote the length of the minimum cost path. Then this random variable satisfies the following asymptotics:*

$$n\mathbf{L}^{\mathrm{unif}}_{12} - \log n \overset{d}{\to} \Xi^1_{W_1 \cdot W_2}$$

as $n \to \infty$. Here as before, W_1, W_2 are independent exponential(1) *random variables.*

Proof: Let L_e be the random variable on edge e with exp(1) distribution, iid over edges. Define the random variable U_e by the relation

$$L_e = -\log(1 - U_e).$$

Note that for any path $\mathcal{P}$ between nodes 1 and 2, the cost of the paths under the two schemes $(U_e)_{e \in E_n}$ and $(L_e)_{e \in E_n}$ satisfies the relation

$$\sum_{e \in \mathcal{P}} U_e \le \sum_{e \in \mathcal{P}} L_e \le \sum_{e \in \mathcal{P}} U_e + \sum_{e \in \mathcal{P}} O(U_e^2) \le \sum_{e \in \mathcal{P}} U_e + O\left(\left(\sum_{e \in \mathcal{P}} U_e\right)^2\right). \tag{22}$$

In this coupling write $\mathbf{L}^{\mathrm{unif}}_{12}$ and $\mathbf{L}^{\mathrm{exp}}_{12}$ to denote the length of the shortest path under the uniform (U_e) environment and exponential (L_e) environment. Since the shortest path under the exponential distribution is of order $\log n / n$, the above relations imply that the shortest path under uniform edge weights $\mathbf{L}^{\mathrm{unif}}_{12}$ satisfies the relation $L^{\mathrm{unif}}_{12} = O_P(\log n / n)$. In particular, Eq. (22) implies that

$$\mathbf{L}^{\mathrm{unif}}_{12} \le \mathbf{L}^{\mathrm{exp}}_{12} \le \mathbf{L}^{\mathrm{unif}}_{12} + O((\mathbf{L}^{\mathrm{unif}}_{12})^2) \le \mathbf{L}^{\mathrm{unif}}_{12} + O_P\left(\frac{\log n}{n}\right)\mathbf{L}^{\mathrm{unif}}_{12}.$$

Thus

$$nO_P\left(\frac{\log n}{n}\right)L_{12}^{\text{unif}} = O_P\left(\frac{\log^2 n}{n}\right) \to 0$$

as $n \to \infty$. Now by the proven asymptotics for $\mathbf{L}_{12}^{\text{exp}}$ the result for $\mathbf{L}_{12}^{\text{unif}}$ follows as well. $\qquad\square$

Proof of Theorem 2 part (b): Recall the definition of a random recursive tree $\mathcal{T}_k$ on k nodes from Sec. II. We first quote some well known asymptotic properties of this random tree. It is taken from the quite comprehensive survey in Ref. 11.

Proposition 16 (Ref. 11): *Let V_k be a node chosen uniformly at random from the tree $\mathcal{T}_k$, and let D_k denote the number of edges on the path from the root to this node. Then*

$$\frac{D_k}{\log k} \to_P 1.$$

Consider the simultaneous flows from nodes 1 and 2 as constructed above. For times t before the collision time S_{12}, let $\mathcal{T}^{(i)}(t)$ denote the tree consisting of all shortest paths from node i to nodes in $\mathcal{F}_i(t)$ for $i = 1, 2$. Then the following lemma ties up these shortest path trees with the random recursive tree.

Lemma 17: *The random trees $\mathcal{T}^{(1)}(S_{12})$ and $\mathcal{T}^{(2)}(S_{12})$ are distributed as random recursive trees on $N_n^{(1)}(S_{12})$ and $N_n^{(2)}(S_{12})$ nodes, conditionally independent of each other given the random variables $N_n^{(i)}(S_{12})$. Further the collision is made by joining up a uniformly random node in $\mathcal{T}^{(1)} \times (S_{12})$ with a uniformly random node in $\mathcal{T}^{(1)}(S_{12})$.*

Proof: By the symmetry of the flow processes on the complete graph, note that before collision, any node v that is seen by the flow from 1 at some time t is seen via an edge from some attached uniformly at random to one of the nodes already present in the flow cluster $\mathcal{F}^{(1)}(\cdot)$ before time t. A similar argument holds for the flow emanating from node 2 and this completes the proof. $\qquad\square$

Thus Lemma 17 implies that we have the distributional identity

$$H_n = D_{N_n^{(1)}(S_{12})}^1 + D_{N_n^{(2)}(S_{12})}^2 + 1. \tag{23}$$

Note that part (a) of this theorem implies that $S_{12} = \frac{1}{2}\log n + O_P(1)$. In particular, Proposition 13 on the rate of growth of the two random variables $N_n^{(i)}(t)$ for $i = 1, 2$ implies that given any $\varepsilon > 0$, we can choose constants $0 < A, B < \infty$ (depending on ε) such that

$$\liminf_{n \to \infty} \mathbb{P}(A\sqrt{n} \le N_n^{(1)}(S_{12}) \le B\sqrt{n}, \ A\sqrt{n} \le N_n^{(2)}(S_{12}) \le B\sqrt{n}) \ge 1 - \varepsilon.$$

Thus by Lemma 17 we have

$$\frac{D_{N_n^{(i)}(S_{12})}^i}{\frac{1}{2}\log n} \to_P 1$$

as $n \to \infty$.

Now use Eq. (23) to conclude that

$$\frac{H_n}{\log n} \to_P 1$$

as $n \to \infty$. $\qquad\square$

Proof of Theorem 2 part (c): This part essentially uses the characterization given by Eq. (14) and large deviation estimates for the exponential distribution. For this we need to understand the exact dynamics of the growth of the shortest path tree from node 1 on the complete graph with exponential edge weights. We shall as before assume that we have rescaled all edge lengths so that all the edges have an exponential distribution with mean n. Then note that we need to show

$$\mathbb{P}(\exists e \in \mathcal{T}_1, L_e > a \log n) \to 0$$

as $n \to \infty$. Let $(Y_i)_{i \geq 1}$ be a sequence of independent random variables distributed as exponential random variables with mean 1. Then the tree is constructed by sequentially adjoining $(n-1)$ edges as follows.

(1)   At time 0 start with a single node labeled 1.
(2)   At time $T_1 = [n/(n-1)]Y_i$ attach a node labeled 2 to this node.
(3)   Recursively proceed as follows: after attaching the first k nodes, attach the $(k+1)$th node at time $T_k = T_{k-1} + [n/k(n-k)]Y_k$ uniformly at random to one of the nodes $1, 2, \ldots, k$. If the node is attached to node $m \leq k$ then this edge has length $T_k - T_m$.
(4)   Continue for $k \leq n-1$.

This construction, in particular, implies that the length of the kth edge, L_k, has the distribution

$$L_k = T_k - T_m \quad \text{with probability} \ 1/k. \tag{24}$$

Now we need to show that

$$\sum_{k=1}^{n-1} \mathbb{P}(L_k > a \log n) \to 0$$

as $n \to \infty$. Note that

$$\mathbb{P}(L_k > a \log n) = \mathbb{P}(k+1 \ \text{attached to} \ 1, L_k > a \log n) + \mathbb{P}(k+1 \ \textbf{not} \ \text{attached to} \ 1,$$

$$L_k > a \log n).$$

Fix $\varepsilon > 0$. Since for $k \geq n(1-\varepsilon)$ we have

$$\mathbb{P}(k+1 \ \text{attached to} \ 1, L_k > a \log n) \leq \frac{1}{n(1-\varepsilon)} \tag{25}$$

so that

$$\sum_{k \geq n(1-\varepsilon)} \mathbb{P}(k+1 \ \text{attached to} \ 1, L_k > a \log n) \leq \frac{\varepsilon}{1-\varepsilon},$$

it is thus enough to prove the following for any fixed $\varepsilon > 0$:

(a) $a_k^n := \mathbb{P}(k+1 \ \textbf{not} \ \text{attached to} \ 1, L_k > a \log n) = o(1/n)$ for all k.
(b)   For all $k \leq n(1-\varepsilon)$, $b_k^n = \mathbb{P}(k+1 \ \text{attached to} \ 1, L_k > a \log n) = o(1/n)$.

To prove (a), by Eq. (24) we have

$$a_k^n = \frac{1}{k}\sum_{m=2}^{k-1} \mathbb{P}\left(\sum_{j=m}^{k} \frac{n}{j(n-j)}Y_j > a \log n\right) \leq \frac{1}{k}\sum_{m=2}^{k-1}\left[e^{-\lambda a \log n}\prod_{j=m}^{k}\left(1 - \frac{\lambda n}{j(n-j)}\right)^{-1}\right] \tag{26}$$

by Markov's inequality, where $0 < \lambda < 1 - 1/n$.

Now write

$$\prod_{j=m}^{k}\left(1 - \frac{\lambda n}{j(n-j)}\right)^{-1} = \exp\left[-\sum_{j=m}^{k}\log\left(1 - \frac{n}{j(n-j)}\right)\right] \leq \exp\left(\sum_{j=m}^{k}\frac{n}{j(n-j)} + O\left(\frac{n}{j(n-j)}\right)^2\right).$$

By the summability of the series $\Sigma_1^\infty 1/j^2$, we see that $\Sigma_m^k O((n/j(n-j)))^2) \leq C$. Thus to bound the inequality in Eq. (26), letting $\lambda = 1 - 1/\log n$ and by some simple algebraic manipulations implies that $a_k^n \leq O\left(\frac{\log n}{n^a}\right)$. This proves (a).

To prove (b) note that

$$b_k^n \leq \frac{C}{k} \frac{e^{-\lambda a \log n}}{\left(1 - \dfrac{n\lambda}{n-1}\right)} e^{-\Sigma_2^k(n/j(n-j))}$$

for some constant C independent of ε. Using the fact that $k \leq n(1-\varepsilon)$, it is easy to show that for $\lambda = 1 - 1/\log n$, via some simple algebraic manipulations we get

$$b_k^n \leq \frac{1-\varepsilon}{\varepsilon} \frac{C \log n}{n^a}.$$

This proves the result.   □

B. Multiple source destination pairs

In this section we shall prove Theorem 3. We recall the setting of the above result, namely, we are in the setting of the complete graph with iid exponential edge lengths and as before, for ease of stating results, we shall assume that all edge lengths are iid exponential with mean n instead of mean 1. The basic conceptual idea of the proof is the same as the proof of Theorem 2 part (a), namely, the simultaneous expansion of flow neighborhoods about different nodes. We shall first prove the following result.

Theorem 18: *Fix a finite set of nodes $\{1, 2, \ldots, k\}$ in $\mathcal{G}_n$. Let $W_1, \ldots, W_k$ be independent exponentially distributed random variables with mean 1. Then the random array $(\mathbf{L}_{ij}^1, 1 \leq i < j \leq k)$ satisfies the asymptotics*

$$(\mathbf{L}_{ij}^1 - \log n, 1 \leq i < j \leq k) \overset{d}{\to} (\Xi_{W_i \cdot W_j}^1 : 1 \leq i < j \leq k), \tag{27}$$

where, given the sequence W_i, $\Xi_{W_i \cdot W_j}$ are conditionally independent Poisson point processes on $\mathbb{R}$, with intensity functions given by $f_{ij}(s) = W_i \cdot W_j e^s$, and $\Xi_{W_i \cdot W_j}^1$ is the first point of this point process.

A simple algebraic exercise then allows us to conclude that the distribution of the random variables in Eq. (27) and those in terms of double exponential distributions and logistic distributions as expressed in Eqs. (10) and (11) are identical. Thus it is enough to prove Theorem 18.

Proof: First consider the pruned percolation flow processes as defined in Sec. V A for first passage percolation flows started simultaneously from k different sources. As before we again assume that all the simultaneous percolation flows are **pruned**, namely, if the flow from some node i reaches some node already seen by the flow from some other node j, then this branch of the flow is stopped. As before $N_n^{(i)}(t)$ is used to denote the number of nodes which were first seen by node i by time t.

Now analogous to Proposition 13 we have the following result which is proved almost identically and we skip the proof.

Proposition 19: *Let $\omega_n = o(\log n)$ be a sequence with $\omega_n \to \infty$ and let $B > 0$ be a constant arbitrarily large. Fix k nodes, say, $1, 2, \ldots, k$, and consider first passage percolation flow started simultaneously from these k nodes. Let $\widetilde{\mathcal{F}}(t)$ be the combined filtration of these k flows and let $N_n^{(i)}(t)$ be the number of nodes seen by the pruned percolation flow from node i. Then there exist independent random variables $(W_n^i)_{1 \leq i \leq k}$, exponentially distributed with rate 1 such that for all i, we have*

$$\sup_{\omega_n \leq t \leq (1/2)\log n + B} \left| \frac{N_n^{(i)}(t)}{e^t} - W_n^{(i)} \right| \to_P 0$$

as $n \to \infty$.

Fix $1 \leq i < j \leq k$ and let the random variables $S_{i,j}$ denote the time for the first collision between the first clusters of nodes i and j in the pruned flow processes. Let $\lambda_n^{ij}(\cdot)$ denote the hazard function

$$\lambda_n^{(ij)}(s) = \frac{N_n^{(i)}(s)N_n^{(j)}(s)}{n}.$$

(28)

Using an argument identical to that following Proposition 13 and using Proposition 19 gives the proof of the following proposition.

Proposition 20: *The array $(S_{ij}:1 \leq i < j \leq k)$ satisfies the following asymptotics:*

$$(2S_{ij} - \log n, \ 1 \leq i < j \leq k) \overset{d}{\rightarrow} (\Xi_{W_i \cdot W_j}^1 : 1 \leq i < j \leq k).$$

Now *a priori*, because of our definition of pruned percolation by restricting flow through collisions, it is not true that $2S_{ij}$ and $\mathbf{L}_{ij}^1$ coincide. However, the following deterministic lemma says that on an event (which has high probability for large n), the two in fact do coincide. We restrict to the case where $k=3$; however, there is a similar lemma in the general case which completes the proof. The crucial point is that we need to understand the effect of pruning on the discovery of the shortest paths.

The setting of the deterministic lemma is as follows: Consider a graph $\mathcal{G}$ with deterministic edge lengths $\{l(e) > 0 : e \in \mathcal{G}\}$, such that the shortest path between every pair of nodes is unique. Let L_{ij} denote the length of the shortest path from node i to node j. Fix three nodes, say, 1–3, and start the pruned first passage percolation flow from these three nodes simultaneously. Let S_{ij} be the time when the flow cluster of node i collides with the flow cluster from node j.

Suppose the following conditions hold for the times of collisions and places where collisions take place.

(a)   To fix a definite order as to when the collisions take place, suppose $S_{12} < S_{13} < S_{23}$. Let $\mathbb{F}_i(t)$ denote the **pruned percolation cluster** explored by the flow from node i by time t. Suppose that the first collision takes place via an edge from node u_1 in $\mathbb{F}_1(S_{12})$ with a node v_1 in $\mathbb{F}_2(S_{12})$. Suppose that in the time interval $(S_{12}, S_{13}]$, there occur $k \geq 1$ collisions between these two flow clusters at times $S_{12}^2 < S_{12}^3 < \cdots < S_{12}^k$ via edges $\{(u_2, v_2), (u_3, v_3), \ldots, (u_k, v_k)\}$, where we assume that all the points u_i and v_j are distinct.

(b)   Let $\mathcal{B}(v, r)$ denote the neighborhood of radius r around a node v in $\mathcal{G}$; similarly for a set of nodes A, let $\mathcal{B}(A, r)$ denote the set of all nodes within distance r from the set A. Assume that in $\mathcal{G}$, for the collision nodes $v_i \in \mathbb{F}_2(S_{12}^i)$ defined above, we have the following fact holding: The intersection $\mathcal{B}(v_i, S_{13} - S_{12}^i) \cap \mathcal{B}(\mathbb{F}_3(S_{12}^i), S_{13} - S_{12}^i) = \emptyset$.

(c)   Finally for analyzing the condition for $L_{23} = 2S_{23}$ we have a condition similar to (b) but slightly more complicated because it involves two sets of collisions, one involving collisions between $\mathbb{F}_1(\cdot)$ and $\mathbb{F}_2(\cdot)$ and another set involving collisions between $\mathbb{F}_1(\cdot)$ and $\mathbb{F}_3(\cdot)$. Associated with each flow process $\mathbb{F}_i(\cdot)$, consider the *adjoined flow* processes $\hat{\mathbb{F}}_i$, formed by adjoining to $\mathbb{F}_i(t)$, the nodes belonging to any other flow process $\mathbb{F}_j(t)$ through which a collision occurs with $\mathbb{F}_i(t)$ at some time t. Thus, e.g., if a collision happens between the flow clusters $\mathbb{F}_1(\cdot)$ and $\mathbb{F}_2(\cdot)$ at time t through an edge (u, v) with $u \in \mathbb{F}_1(t)$ and $v \in \mathbb{F}_2(t)$ then $v \in \hat{\mathbb{F}}_1(t)$. Call a new node that belongs to $\mathbb{F}_j(\cdot)$ but is adjoined to the $\mathbb{F}_i(\cdot)$ through a collision at time t, a $j \hookrightarrow i(t)$ adjoined of $\hat{\mathbb{F}}_i(t)$. Then we assume that

(i)   for any $1 \hookrightarrow 2(t)$ collision u originally in $\mathbb{F}_1(t)$, the intersection of the neighborhood $\mathcal{B}(u, S_{23} - t) \cap \mathcal{B}(\hat{\mathbb{F}}_3(t), S_{23} - t) = \emptyset$ and

(ii)   for any $1 \hookrightarrow 3(s)$ collision via node w in $\mathbb{F}_1(s)$, the intersection of the neighborhood of the set $\mathcal{B}(w, S_{23} - s) \cap \mathcal{B}(\hat{\mathbb{F}}_2(s), S_{23} - s) = \emptyset$.

Lemma 21: *Assume that conditions (a)–(c) hold. Write L_{ij} for the length of the shortest path between nodes i and j . Then we have*

$$L_{12} = 2S_{12}, \quad L_{13} = 2S_{13}, \quad L_{23} = 2S_{23}.$$

Proof: Consider the **unpruned flow** processes (where we do **not** stop flow through collision

nodes). Let the collision times between these unpruned flow processes be denoted by $\widetilde{S}_{ij}$ and note that the length of the shortest path between nodes i and j satisfies the identity $L_{ij} = 2\widetilde{S}_{ij}$. Then it is easy to check that conditions (a)–(c) ensure that $S_{ij} = \widetilde{S}_{ij}$ and this completes the proof.   □

Now consider the pruned first passage percolation flow simultaneously from the three nodes 1–3 in the random network $\mathcal{G}_n$. Let A_n be the event that the three conditions (a)–(c) hold for the pruned percolation process. Then the following lemma shows that with high probability the collision times actually coincide with the length of the shortest paths with high probability. Combining with Proposition 20, this completes the proof.   □

Lemma 22: *For the event A_n defined above, we have $\mathbb{P}(A_n) \to 1$ as $n \to \infty$ so that with high probability*

$$2S_{12} = L_{12}, \quad 2S_{13} = L_{13}, \quad 2S_{23} = L_{23}.$$

Proof: Note that for any fixed B

$$\mathbb{P}(A_n) = \mathbb{P}\left(A_n, \max_{1 \le i < j \le 3} S_{ij} < \frac{1}{2}\log n + B\right) + \mathbb{P}\left(A_n, \max_{1 \le i < j \le 3} S_{ij} > \frac{1}{2}\log n + B\right).$$

By Proposition 20, given any $\varepsilon > 0$ we can choose $B = B(\varepsilon)$ such that

$$\limsup_{n \to \infty} \mathbb{P}\left(\max_{1 \le i < j \le 3} S_{ij} > \frac{1}{2}\log n + B\right) \le \varepsilon.$$

Fix this B. We shall now show that for this B

$$\mathbb{P}\left(A_n, \max_{1 \le i < j \le 3} S_{ij} < \frac{1}{2}\log n + B\right) \le o(1)$$

as $n \to \infty$ and this completes the proof. Note that by the definition of the event A_n we have

$$\left\{A_n, \max_{1 \le i < j \le 3} S_{ij} < \frac{1}{2}\log n + B\right\} \subseteq \left\{\exists v \ne \{1,2,3\} : d(v,i) \le \frac{1}{2}\log n + B \quad \forall\, i = 1,2,3\right\},$$

where as before $d(v,i)$ denotes the distance metric induced on the complete graph by the random $\exp(n)$ edge weights. Denote the event on the right by B_n. Then note that by the union bound we have

$$\mathbb{P}(B_n) \le n\mathbb{P}\left(\{1,2,3\} \subseteq \mathcal{B}(n, \tfrac{1}{2}\log n + B)\right), \tag{29}$$

where $\mathcal{B}(n, \tfrac{1}{2}\log n + B)$ is the neighborhood of radius $\tfrac{1}{2}\log n + B$ about the node n. Define the random variable $Y_n = |\mathcal{B}(n, \tfrac{1}{2}\log n + B)|$. Then note that by symmetry, conditional on Y_n, we have

$$\mathbb{P}(B_n | Y_n) = \frac{\dbinom{n-4}{Y_n - 3}}{\dbinom{n-1}{Y_n}}.$$

Thus uniformly on the set $\{Y_n \le 3B\sqrt{n}\, \log n\}$

$$\mathbb{P}(B_n | Y_n) \le (3B)^3 \frac{n^{3/2} \log^3 n}{n^3}.$$

Thus to finish the proof of Eq. (29) it is enough to show

$$\mathbb{P}(Y_n > 3B\sqrt{n}\,\log n) = o\left(\frac{1}{n}\right). \tag{30}$$

Equations (14) and (15) imply that the first passage percolation flow process is dominated by the Yule process and, in particular, $Y_n \leq_{st} N(\frac{1}{2}\log n + B)$ where $N(\cdot)$ is a Yule process. The proof of Eq. (30) now follows because $N(\frac{1}{2}\log n + B) \sim \mathrm{Geom}(\frac{1}{2}\log n + B)$. □

C. Vickrey–Clarke–Groves measure of overpayment

Here we formulate the fundamental heuristic that allows us to do the necessary computations. As before via rescaling, we shall assume that all edge lengths are exponential with mean n (rate $1/n$) random variables.

1. Fundamental identity

Consider the following set of identities which hold for any random graph model which is transitive, i.e., the neighborhood of every node looks exactly the same. Denote the shortest path tree from 1 to all the other nodes as $\mathcal{T}$,

$$
\begin{aligned}
\mathbb{E}_\nu(\mathrm{TB}) &= \mathbb{E}\left(\sum_{e \in \mathcal{T}} [\mathbf{L}_{1,2}(\mathcal{G}_n \setminus e) - \mathbf{L}_{1,2}(\mathcal{G}_n)]\mathbb{1}_{\{e \in \boldsymbol{\pi}(1,2)\}}\right) = \frac{1}{n-1}\mathbb{E}\left(\sum_{k=2}^{n} \sum_{e \in \mathcal{T}} [\mathbf{L}_{1,k}(\mathcal{G}_n \setminus e) - \mathbf{L}_{1,k}(\mathcal{G}_n)]\mathbb{1}\{e\right. \\
&\left. \in \boldsymbol{\pi}(1,k)\}\right) = \frac{1}{n-1}\mathbb{E}\left(\sum_{k=2}^{n} \sum_{e \in \mathcal{T}} \mathbf{L}_{1,k}(\mathcal{G}_n \setminus e)\mathbb{1}\{e \in \boldsymbol{\pi}(1,k)\} - \sum_{k=2}^{n} \sum_{e \in \mathcal{T}} \mathbf{L}_{1,k}(\mathcal{G}_n)\mathbb{1}\{e \in \boldsymbol{\pi}(1,k)\}\right) \\
&= \frac{1}{n-1}\mathbb{E}\left(\sum_{k=2}^{n} \sum_{e \in \mathcal{T}} \mathbf{L}_{1,k}(\mathcal{G}_n \setminus e)(1 - \mathbb{1}\{e \notin \boldsymbol{\pi}(1,k)\}) - \sum_{k=2}^{n} \sum_{e \in \mathcal{T}} \mathbf{L}_{1,k}(\mathcal{G}_n)\mathbb{1}\{e \in \boldsymbol{\pi}(1,k)\}\right) \\
&= \frac{1}{n-1}(\mathcal{A}_1 - \mathcal{A}_2), \tag{31}
\end{aligned}
$$

where

$$\mathcal{A}_1 = \mathbb{E}\left(\sum_{k=2}^{n} \sum_{e \in \mathcal{T}} \mathbf{L}_{1,k}(\mathcal{G}_n \setminus e)\right)$$

and

$$\mathcal{A}_2 = \mathbb{E}\left(\sum_{k=2}^{n} \sum_{e \in \mathcal{T}} \mathbf{L}_{1,k}(\mathcal{G}_n)\right) = (n-1)\mathbb{E}\left(\sum_{k=2}^{2} \mathbf{L}_{1,k}(\mathcal{G}_n)\right) \text{ since there are } n-1 \text{ edges in}$$

$$\mathcal{T} = (n-1)^2 \gamma_n(\mathcal{G}_n),$$

where $\gamma_n(\mathcal{G}_n) = \mathbb{E}(\mathbf{L}_{1,2}(\mathcal{G}_n))$ is the expected cost of the least cost path between the two nodes 1 and 2 and by Theorem 2 part (a), $\gamma_n = \log n + o(1)$ as $n \to \infty$. Also note that here we have used the identity

$$\sum_{k=2}^{n} \sum_{e \in \mathcal{T}} \mathbf{L}_{1,k}(\mathcal{G}_n \setminus e)\mathbb{1}\{e \notin \boldsymbol{\pi}(1,k)\} = \sum_{k=2}^{n} \sum_{e \in \mathcal{T}} \mathbf{L}_{1,k}(\mathcal{G}_n)\mathbb{1}\{e \notin \boldsymbol{\pi}(1,k)\}.$$

Thus $\rho(\mathcal{G}_n)$ can be written as

$$\rho(\mathcal{G}_n) = \frac{\mathcal{A}_1}{(n-1)^2 \gamma_n(\mathcal{G}_n)} - 1. \tag{32}$$

Note that the above heuristic works for general random graphs which are "homogeneous" in the sense that the neighborhood of every node statistically looks the same, e.g., the configuration model and Erdős–Rényi random graphs.

So the above set of equations suggests the following computation for "homogeneous graphs."

(a)   First passage percolation: Compute the expected amount of time to get from node 1 to node 2, namely, $\gamma_n(\mathcal{G}_n) = \mathbb{E}(\mathbf{L}_{1,2}(\mathcal{G}_n))$ for uniform random choices of source destination pair (s,t). This gives us the denominator for $\rho(\mathcal{G}_n)$.

(b)   Numerator:

$$\mathcal{A}_1 = \mathbb{E}\left(\sum_{k=2}^{n} \sum_{e \in \mathcal{T}} \mathbf{L}_{1,k}(\mathcal{G}_n \setminus e)\right).$$

We shall show that in many cases this term is also easily computable.

2. Point to point distances associated with VCG on complete graph

To get the conceptual idea of the computations involved first consider the following separate proof of Theorem 2. This is how the result was proved in Ref. 6. Start a flow from node 1 which percolates through the graph at unit rate. Let v_k be the kth node seen by the flow. Let T_k be the time taken by the flow to see this node. Finally let $Z_k = T_k - T_{k-1}$ so that $T_k = \Sigma_{i=1}^{k} Z_i$. The following lemma (essentially from Ref. 6) gives the rules for recursively reconstructing these nodes and times.

Lemma 23:

(i)   Let node $v_0 = 1$ and $T_0 = 0$. Then inductively on k , conditional on $v_0, \ldots, v_{k-1}$ v_k and Z_k are independent with

$$v_k \sim \mathrm{unif}([n] \setminus v_0, v_1, \ldots, v_{k-1}),$$

$$Z_k \sim \exp\left(\frac{n}{k(n-k)}\right).$$

(ii)   The above relations easily yield that $\gamma(\mathcal{G}_n) \sim \log n$ as $n \to \infty$.

Now note that

$$\mathcal{A}_1 = \sum_{k=1}^{n-1} \mathbb{E}\left(\sum_{l=1}^{n-1} \mathbf{L}_{1,v_l}(\mathcal{G}_n \setminus e_k)\right),$$

where e_k denotes the edge $e \in \mathcal{T}$ that attached v_k to $\mathcal{T}$. The following crucial lemma collects the facts that allow us to analyze what happens to the cost of paths when we delete e_k. Compare with Lemma 23.

Fix $k \geq 1$. Delete edge e_k from $\mathcal{T}$. Start the flow as before from node 1. Let $\hat{v}_j$ denote the jth node seen by the flow in this new $\mathcal{G}_n \setminus e_k$ and let $\hat{T}_k$ be the time to see node $\hat{v}_k$. As before let $\hat{Z}_j = \hat{T}_j - \hat{T}_{j-1}$ so that $T_j = \Sigma_1^{j} \hat{Z}_j$. Note that $\mathbb{E}(\Sigma_{l=1}^{n-1} \mathbf{L}_{1,v_l}(\mathcal{G}_n \setminus e_k)) = \mathbb{E}(\Sigma_{j=1}^{n-1} \hat{T}_j)$.

Lemma 24: Fix $k \geq 1$. The following identities hold.

(i)   For $j < k$, $\hat{v}_j = v_j$, $\hat{Z}_j = Z_j$.
(ii) $\hat{Z}_k \sim \hat{Y}_k[n/k(n-k)] + Y_k(n/[k(n-k)-1])$.
(iii) For $j \geq k$, conditional on $\hat{v}_0, \ldots, \hat{v}_{j-1}$, we have the following distributional identities.

(a)   *If $v_k \in \{\hat{v}_0, \ldots, \hat{v}_{j-1}\}$ then*

$$\hat{v}_j \sim \mathrm{unif}([n] \setminus v_0, \ldots, v_{j-1}),$$

$$Z_j \sim \frac{n}{j(n-j)} Y_j,$$

and Z_j and $\hat{v}_j$ are conditionally independent.

(b)   *If $v_k \notin \{\hat{v}_0, \ldots, \hat{v}_{j-1}\}$ then $\hat{Z}_j \sim [n/(j(n-j)-1)]Y_j$ and $\hat{v}_j$ and $\hat{Z}_j$ are conditionally independent. The (conditional) probability that $\hat{v}_j$ is equal to v_k is $(j-1)/[j(n-j)-1]$. Finally conditional on $\hat{v}_j \neq v_k$, $\hat{v}_j \sim \mathrm{unif}([n]\setminus\{\hat{v}_0, \ldots \hat{v}_{j-1}, v_k\})$.*

Completing the proof of Theorem 4: Using the relations in Lemma 24, after some algebra, the following is easily given:

$$\mathbb{E}\left(\sum_{l=1}^{n-1} \mathbf{L}_{1,\hat{v}_l}(\mathcal{G}_n \setminus e_k)\right) = (n-1)\gamma(\mathcal{G}_n) + \frac{n}{k} + C_k,$$

where $C_k \leq \pi^2/6 + 1/k^2$. Summing over k gives us the result.

D. Proofs for the dense Erdős–Rényi random graph

The proofs essentially follow via a coupling with the setup on the complete graph and the deletion of "large edges."

Proof of Theorem 5 part (a): Consider the complete graph with $\exp(1)$ edge weights. For simplicity we shall assume instead of Eq. (12) the following condition, namely, that there exists an $1 < a \leq \infty$ such that for all n large enough,

$$p_n > \frac{a \log n}{n}. \tag{33}$$

Let l_n be a sequence solving the equation

$$p_n = P(X \leq l_n) = 1 - e^{-l_n},$$

where $X \sim \exp(1)$. Note that because of condition (33), for any $\varepsilon > 0$, there exists $N = N_\varepsilon$ large enough such that for all $n \geq N$,

$$l_n > \frac{(1-\varepsilon)a \log n}{n}. \tag{34}$$

Choose $\varepsilon > 0$ small enough so that $a(1-\varepsilon) > 1$. Let $\mathcal{T}_1$ denote the shortest path tree from node 1 to all other nodes. Note that the Erdős–Rényi random graph $\mathcal{G}_n^{p_n} = \mathrm{ER}(n, a \log n/n)$ can be constructed by deleting all edges from the complete graph which have edge length greater than l_n. Theorem 2 part (d) [with a replaced by $a(1-\varepsilon)$] and Eq. (34) implies that in this construction, $\mathcal{T}_1 \subseteq \mathcal{G}_n^{p_n}$ with high probability. Since $\mathcal{T}_1$ is a connected graph spanning the node set, this implies that $\mathcal{G}_n^{p_n}$ is connected with high probability. $\square$

Proof of Theorem 5 part (b): This follows almost trivially by Proposition 15. First note that we can construct $\mathcal{G}_n^{p_n}$ with uniform edge weight distributions as follows.

(a)   Consider the complete graph with $U[0,1]$ edge weights.
(b)   Delete all edges which have length greater that p_n. This gives the Erdős–Rényi random graph where all the edge weights are uniformly distributed on $[0, p_n]$.
(c)   Let X_e be the length of any edge e currently in the graph. Let $Y_e = (1/p_n)L_e$. This gives an Erdős–Rényi random graph where all the edge lengths are $U[0,1]$ random variables.

Let $\mathbf{L}_{12}^{\text{comp}}$ be the length of the shortest path in the complete graph and let $\mathbf{L}_{12}^{\text{er}}$ be the length of the shortest path in the above construction in the Erdős–Rényi random graph, setting $\mathbf{L}_{12}^{\text{ER}}=0$ if the two nodes are disconnected. Note that Proposition 15 implies, in particular, that for the complete graph with uniform edge weights

$$\frac{\mathbf{L}_{12}^{\text{comp}}}{\log n/n} \to_P 1$$

as $n \to \infty$. In particular, this implies that in the above construction,

(i)   the two nodes 1 and 2 are connected with high probability and
(ii)   in the above construction $\mathbf{L}_{12}^{\text{Er}}=(1/p_n)\mathbf{L}_{12}^{\text{comp}}$ and more pertinently the shortest paths in the two graphs, namely, the complete graphs $\mathcal{G}_n$ and $\mathcal{G}_n^{p_n}$, are identical.

The result then follows from asymptotics for $\mathbf{L}_{12}^{\text{comp}}$, namely, Proposition 15.   □

Proof of Theorem 5 part (c): Suppose we know that the hopcount for the complete graph with uniform edge weights between nodes 1 and 2, namely, H_n^{comp}, is $\approx \log n$. Then by the above coupled construction of $\mathcal{G}_n^{p_n}$ and observation (ii) above that in the coupled construction, the shortest paths coincide in the two models, we would have that the hopcount in the Erdős–Rényi random graph $H_n^{\text{Er}} \approx \log n$. The following behavior of the hopcount in the complete graph with uniform edge weights was proved by Janson, see Ref. 6.

Proposition 25 (Ref. 6): *Consider the complete graph with* $U[0,1]$ *edge weights. Then the hopcount of the shortest path,* H_n^{comp}, *has the asymptotics*

$$\frac{H_n^{\text{comp}}}{\log n} \to_P 1.$$

This gives our result.   □

Proof of Theorem 7: We shall now show how Theorem 7 essentially follows from the proof of Theorem 5, coupled with some additional facts about the behavior of the gamma distribution near 0. We first briefly describe the proof technique. We first prove the behavior of the hopcount of the optimal path between nodes 1 and 2 and then prove asymptotic distributional limits for the actual length of the path. Let $\mathcal{G}_n^{p_n}$ be the Erdős–Rényi random graph with exp(1) edge weights. Via a coupling procedure with the $U[0,1]$ case we shall first show that, assuming Eq. (13) holds, there exists a path between 1 and 2 of length $\log n + O_P(1)/np_n$ with the number of edges on this path approximately $\log n$. Then via the properties of the gamma distribution we shall show that for any fixed $\varepsilon > 0$, all paths between nodes 1 and 2, with either $(1-\varepsilon)\log n$ edges or $(1+\varepsilon)\log n$ edges, have to have length much larger than $(\log n + B)/np_n$ for any fixed p_n and any fixed constant $B > 0$. Combining this with the previous fact implies that the number of edges on shortest path between nodes 1 and 2 has to $\in [(1-\varepsilon)\log n, (1+\varepsilon)\log n]$ with high probability as $n \to \infty$. Since ε was arbitrary, this completes the proof for the hopcount. We now formalize the above ideas.

Proposition 26: *Consider the Erdős–Rényi random graph* $\mathcal{G}_n^{p_n}$ *with* exp(1) *edge weights, with* p_n *satisfying Eq. (13). Then with probability converging to 1 as* $n \to \infty$, *there exists a path* $\mathcal{P}^*$ *between 1 and 2 satisfying the following two properties.*

(a)   *Let* $\mathbf{L}_{12}^*$ *be the length of this path. Then*

$$np_n\mathbf{L}_{12}^* - \log n \xrightarrow{d} \Xi_{W_1 \cdot W_2}$$

as $n \to \infty$.

(b)   *Let* h_n *be the number of edges on this path. Then*

$$\frac{h_n}{\log n} \to_P 1$$

as $n \to \infty$.

Proof: Consider the Erdős–Rényi random graph with uniform edge weights $U[0,1]$, U_e. Note that by Theorem 5, nodes 1 and 2 are connected and the optimal path $\mathcal{P}^*$ between these two nodes satisfies the following.

(a)   The length $\mathbf{L}_{12}^{\text{unif}}$ satisfies the asymptotics

$$np_n\mathbf{L}_{12}^{\text{unif}} - \log n \xrightarrow{d} \Xi_{W_1 \cdot W_2}.$$

(b)   The hopcount of this path satisfies

$$\frac{h_n}{\log n} \to_P 1$$

as $n \to \infty$.

Now for every edge in $\mathcal{G}_n^{p_n}$ with edge weight U_e define the transformation

$$L_e = -\log(1 - U_e).$$

This gives us the Erdős–Rényi random graph with iid exp(1) edge weights. Let $\mathbf{L}_{12}^*$ be defined as the cost of the path $\mathcal{P}^*$ under this new edge weight structure, namely,

$$\mathbf{L}_{12}^* = \sum_{e \in \mathcal{P}^*} L_e = \sum_{e \in \mathcal{P}^*} -\log(1 - U_e).$$

To finish the proof of Proposition 26, it is enough to show that

$$\mathbf{L}_{12}^* = \mathbf{L}_{12}^{\text{unif}} + R_n,$$

where

$$np_n R_n \to_P 0$$

as $n \to \infty$.

Now note that by Taylor's approximation for the logarithmic function, we have

$$\mathbf{L}_{12}^* = \sum_{e \in \mathcal{P}^*} U_e + \sum_{e \in \mathcal{P}^*} O(U_e^2) = L_{12}^{\text{unif}} + R_n,$$

where the remainder term $R_n \le c_n \mathbf{L}_{12}^*$ where $c_n = \max\{U_e : e \in \mathcal{P}^*\}$. Note that by Theorem 5 since $\mathbf{L}_{12}^* = O_P(\log n/np_n)$, $c_n \le O_P(\log n/np_n)$. In particular, this implies that

$$np_n R_n = O_P\left(\frac{\log^2 n}{n}\right) \to 0$$

as $n \to \infty$, by condition given by Eq. (13). This completes the proof.   □

Proof of Theorem 7 part (a): Note that Proposition 26 implies that in searching for the shortest path between nodes 1 and 2, we can restrict our search to paths which have length less than or equal to $O(\log n/np_n)$. Now note that by the coupling construction of both the uniform and exponential edge weights on the same graph via the transformation $-\log(1-u)$ as used in the proof of Proposition 26, we have for any path $\mathcal{P}$ between nodes 1 and 2 satisfying $\Sigma_{e \in \mathcal{P}} U_e = O_P(\log n/np_n)$

$$\sum_{e \in \mathcal{P}} U_e \le \sum_e L_e \le \sum_{e \in \mathcal{P}} U_e + \sum_{e \in \mathcal{P}} O(U_e^2).$$

We note that the conditions on the lengths of the paths being considered now imply

$$\sum_{e \in \mathcal{P}} O(U_e^2) \le O_P\!\left(\frac{\log n}{np_n}\right) \sum_{e \in \mathcal{P}} U_e.$$

This, in particular, implies that in this coupling of the uniform and exponential edge weights, we have that the respective least cost paths satisfy

$$\mathbf{L}_{12}^{\text{unif}} \le \mathbf{L}_{12}^{\text{exp}} \le \mathbf{L}_{12}^{\text{unif}} + O\!\left(\frac{\log n}{np_n}\right)\mathbf{L}_{12}^{\text{unif}}.$$

Now use Theorem 5 to conclude that

$$np_n\mathbf{L}_{12}^{\text{unif}} - \log n \xrightarrow{d} \Xi_{W_1 \cdot W_2}^1,$$

while for p_n satisfying the condition given by Eq. (13),

$$np_n O\!\left(\frac{\log n}{np_n}\right)\mathbf{L}_{12}^{\text{unif}} = O_P\!\left(\frac{\log^2 n}{np_n}\right) \to_P 0.$$

This implies that

$$np_n\mathbf{L}_{12}^{\text{exp}} - \log n \to_P \Xi_{W_1 \cdot W_2}^1.$$

This completes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Proof of Theorem 7 part (b):

Proposition 27: *Fix any $B>0$ and $\varepsilon>0$ and small. Let $A_{n,\varepsilon}^1$ be the event that there exists a path between nodes 1 and 2 with number of edges **less** than $(1-\varepsilon)\log n$ and actual length less than $(\log n + B)/np_n$. Let $A_{n,\varepsilon}^2$ be the event that there exists a path between nodes 1 and 2 with number of edges **greater** than $(1+\varepsilon)\log n$ and actual length less than $(\log n + B)/np_n$. Then we have*

$$\mathbb{P}(A_{n\varepsilon}^i) \to 0 \quad \text{as } n \to \infty.$$

Proof: We shall show that $\mathbb{P}(A_{n,\varepsilon}^1)$ goes to zero for large n. The proof for $\mathbb{P}(A_{n,\varepsilon}^2)$ is similar and omitted. Fix any $1 \le k \le (1-\varepsilon)\log n$. Fix any path of length k (namely, k consecutive edges starting at 1 and ending at 2 and passing through $k-1$ distinct nodes in between) between nodes 1 and 2. The number of such paths is equal to $\Pi_2^k(n-j)$ and the probability that any such path exists in $\mathcal{G}_n^{p_n}$ is equal to p_n^k. Thus the expected number of paths with k edges which have length less than $(\log n + B)/np_n$ equals

$$\prod_2^k (n-j)p_n^k\mathbb{P}\!\left(\sum_1^k Y_i < \frac{\log n + B}{np_n}\right),$$

where Y_i are independent exp(1) random variables. To bound the probability occurring in the above inequality we quote the following Lemma from Ref. 3 (Lemma 3.3).

Lemma 28: *Let $Y_1, Y_2, \ldots, Y_k$ be k independent exponential random variables with mean 1 and let $u \in [0,1]$. Then*

$$\mathbb{P}\!\left(\sum_1^k Y_i \le u\right) \le 3e^{-u}\frac{u^k}{k!}.$$

Now by Markov's inequality

$$\mathbb{P}(A_{n,\varepsilon}^1) \le \sum_{k=1}^{(1-\varepsilon)\log n}\left[\prod_{j=2}^k (n-j)p_n^k\mathbb{P}\!\left(\sum_1^k Y_i < \frac{\log n + B}{np_n}\right)\right].$$

Now using Lemma 28 we get

$$\mathbb{P}(A_{n,\varepsilon}^1) \le \frac{3}{n} \sum_{1}^{(1-\varepsilon)\log n} \frac{(\log n + B)^k}{k!} = 3e^B \mathbb{P}(\text{Poi}(\log n + B) \le (1-\varepsilon)\log n),$$

where $\text{Poi}(\log n + B)$ is a Poisson random variable with mean $\log n + B$. The result then follows from large deviation principles for the Poisson random variable.   □

To complete the proof of Theorem 7 part (b), given any $\varepsilon > 0$, by Proposition 26, we can choose a $B_\varepsilon > 0$ large such that, writing $\mathbf{L}_{12}^1$ as the shortest path between nodes 1 and 2,

$$\limsup_{n \to \infty} \mathbb{P}\left(\mathbf{L}_{12}^1 > \frac{\log n + B_\varepsilon}{np_n} \right) \le \varepsilon. \tag{35}$$

Proposition 35 with $B = B_\varepsilon$ then implies that

$$\liminf_{n \to \infty} \mathbb{P}(\text{number of edges on optimal path} \in ((1-\varepsilon)\log n, (1+\varepsilon)\log n)) \ge 1 - \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this proves the required asymptotics for the hopcount.   □

VI. CONCLUSION

We see that using simple ideas from continuous time branching processes, we have arrived at a large number of results in a number of different models for the behavior of short paths. We see that the introduction of edge disorder significantly changes the topology of the graph in many cases and increases the transit times (measured in terms of hopcount or number of edges traversed on the minimum cost path) even for ultrasmall networks from $o(\log n)$ to something much larger, approximately $\log n$. We finish with some concluding remarks, including limitations of the above methods and further open problems and conjectures.

(a) Finite n error bounds: Using the previous methods it is probably difficult to arrive at finite n error bounds. See Ref. 14 where using more combinatorial arguments and the inclusion exclusion principle, the authors arrived at finite n error bounds.

(b) We originally got interested in this study because of the fascinating statistical physics paper in Ref. 2 which via simulations predicted that in a wide array of random graph models, even if the original graph distance between typical nodes is sublogarithmic (ultrasmall world graphs) attaching independent edge weights (*weak disorder phase*) tends to change the structure of the graph drastically. The paper in Ref. 2 gives further conjectures regarding what should happen in the setup where the edge weights are such that the maximal edge weight on the path essentially determines the length of the path (*the strong disorder phase*) which is akin to studying the minimal spanning tree in these graphs.

(c) Simultaneous flow processes were also used in Ref. 1 to explore edge flows in the complete graph with exponential edge weights. The setting is slightly more complicated and results in the concept of *size biasing* of the percolation flow clusters.

(d) We state the following conjecture regarding the structure of the shortest path structure between two fixed nodes. In many situations, it is important to know not only the least cost path between two nodes but also the cost second cheapest, third cheapest, and in general the k-cheapest path between two nodes. See Refs. 8 and 4 and the references therein to state just two sources. In the context of this problem, for the complete graph with exponential edge costs, the following conjecture seems plausible and doable.

Conjecture 29: *Let* $\mathbf{L}^i$ *denote the cost of the ith least cost simple path between nodes* 1 *and* 2 *in the complete graph with* $\exp(1)$ *edge costs. Consider the sequence of point process on* $\mathbb{R}$ *given by*

$$\Xi_n = (n\mathbf{L}_{12}^1 - \log n, n\mathbf{L}_{12}^2 - \log n, n\mathbf{L}_{12}^3 - \log n, \ldots).$$

Then

$$\Xi_n \overset{d}{\rightarrow} \Xi_{W_1 \cdot W_2}, \tag{36}$$

where $\overset{d}{\rightarrow}$ denotes weak convergence in the vague topology on measures on $\mathbb{R}$ and W_1, W_2 are independent exponential random variables.

ACKNOWLEDGMENTS

[1] Aldous, D. J. and Bhamidi, S., "Edge flows in the complete random-lengths network," e-print arXiv:0708.0555. ; Random Struct. Algorithms (to appear).

[2] Braunstein, L. A., Buldyrev, S. V., Cohen, R., Havlin, S., and Eugene Stanley, H., "Optimal paths in disordered complex networks," Phys. Rev. Lett. **91**, 168701 (2003).

[3] Fill, J. A. and Pemantle, R., "Percolation, first-passage percolation and covering times for Richardson's model on the n-cube," Ann. Appl. Probab. **3**, 593 (1993).

[4] Hershberger, J., Maxel, M., and Suri, S., Proceedings of the Fifth Workshop Algorithm Engineering and Experiments (ALENEX) (unpublished).

[5] Douglas Howard, C., *Probability on Discrete Structures*, Encyclopedia of Mathematical Science Vol. 110 (Springer, Berlin, 2004), pp. 125–173.

[6] Janson, S., "One, two and three times log n/n for paths in a complete graph with random weights," Combinatorics, Probab. Comput. **8**, 347 (1999); *Random Graphs and Combinatorial Structures* (Mathematisches Forschungsinstitut Oberwolfach, Oberwolfach, 1997).

[7] Karger, D. and Nikolova, E., DIMACS Workshop on Computational Issues in Auction Design, 2004 (unpublished).

[8] Katoh, N., Ibaraki, T., and Mine, H., "Efficient algorithm for K shortest simple paths," Networks **12**, 411 (1982).

[9] Nisan, N. and Ronen, A., in *STOC '99: Proceedings of the 31st Annual ACM Symposium on Theory of Computing*, Atlanta, Georgia (ACM, New York, 1999).

[10] Norris, J. R., *Markov Chains*, Cambridge Series in Statistical and Probabilistic Mathematics Vol. 2 (Cambridge University Press, Cambridge, 1998).

[11] Smythe, R. T. and Mahmoud, H. M., "A survey of recursive trees," Itogi Nauki Tekh. Ser. Teor. Veroyatn. Mat. Stat. Teor. Kibern. **51**, 1 (1994).

[12] Smythe, R. T. and Wierman, J. C., "Percolation on the square lattice," Lecture Notes in Mathematics Vol. 671 (Springer-Verlag, Berlin, 1978).

[13] van den Esker, H., van der Hofstad, R., Hooghiemstra, G., and Znamenski, D., "Distances in random graphs with infinite mean degrees," Extremes **8**, 111 (2005).

[14] van der Hofstad, R., Hooghiemstra, G., and Van Mieghem, P., "First-passage percolation on the random graph," Probab. Eng. Inform. Sci. **15**, 225 (2001).

[15] van der Hofstad, R., Hooghiemstra, G., and Van Mieghem, P., "The flooding time in random graphs," Extremes **5**, 111 (2003).

[16] van der Hofstad, R., Hooghiemstra, G., and Van Mieghem, P., "Distances in random graphs with finite variance degrees," Random Struct. Algorithms **27**, 76 (2005).

[17] van der Hofstad, R., Hooghiemstra, G., and Znamenski, D., "Distances in random graphs with finite mean and infinite variance degrees," Electron. J. Probab. **12**, 703 (2007).

[18] Wastlund, J., Proceedings of the Fourth Colloquium on Mathematics and Computer Science, Algorithms, Trees, Combinatorics and Probabilities, Institut lie Cartan, Nancy, France, 2006 (unpublished).