Journal of Statistical Physics manuscript No.
(will be inserted by the editor)

A Framework for Imperfectly Observed Networks

David Aldous · Xiang Li

Received: date / Accepted: date

Abstract Model a network as an edge-weighted graph, where the (unknown) weight w_e of edge e indicates the frequency of observed interactions, and over time t we observe a Poisson(tw_e) number of interactions across edges e. How should we estimate some given statistic of the underlying network? This leads to wide-ranging and challenging problems, on which this article makes only partial progress.

Keywords network · statistical estimation · community · incomplete

Mathematics Subject Classification (2010) 60J27 · 94C99 · 05C82 · 91D30

1 Introduction

Network science has many aspects: here are two.

Efficient algorithms/computational complexity. Given some mathematically-defined quantity $\Gamma(G)$ associated with a network G, find an algorithm which inputs G and outputs $\Gamma(G)$. Compare different algorithms via theoretical bounds or by contests with real-world network data.

Analysis of probability models. Take a probability model for networks and analyze mathematically some graph-theoretic quantity (degree distribution, diameter, clustering statistics). Or study some random process (e.g. random walk or voter model or Prisoners' Dilemma) over a deterministic network G.

Aldous's research supported by N.S.F Grant DMS-1504802.

David Aldous
Department of Statistics, 367 Evans Hall # 3860, U.C. Berkeley CA 94720;
E-mail: aldous@stat.berkeley.edu

Xiang Li
Department of Statistics, 367 Evans Hall # 3860, U.C. Berkeley CA 94720;
E-mail: lapis.lazuli.8@gmail.com

In the latter context, the expectation of some quantity associated with the process is a functional $\Gamma(G)$.

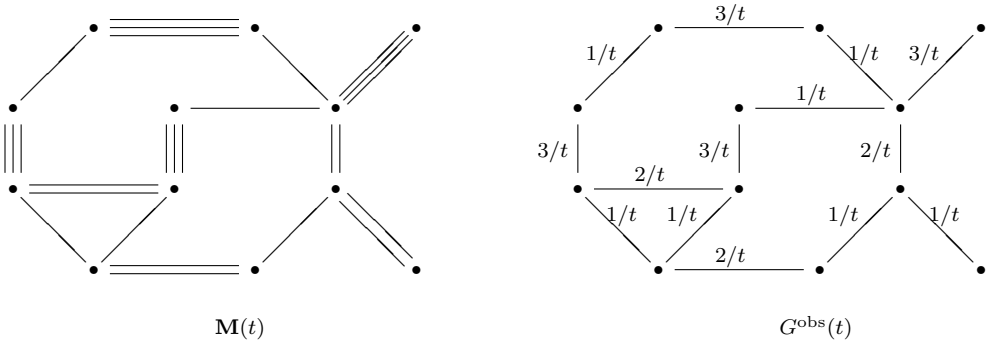
For this article, suppose we are interested in some quantitative question about a real-world network which we could answer if we knew the network. That is, there is some unknown G^{true}, some observed G^{obs} and we want an estimate of $\Gamma(G^{\text{true}})$ for some given functional Γ, and some indication of how accurate the estimate might be. There are many ways to formalize this *imperfectly-observed networks* setting (see section 6.1 for brief comments) motivated by different real-world instances. This article describes a novel framework within which some interesting and challenging mathematical questions arise, though we do not claim any particular real-world relevance. Our framework is rather intermediate between the two aspects above: G^{true} is arbitrary, but we make a probability model for how G^{obs} depends on G^{true}. Also, and important to keep in mind, our implicit notion of "cost" will be observation time – the cost of acquiring data – rather than cost of computation, which we ignore. So this contrasts with a complementary framework called *smoothed analysis* [9] which measures cost of computation of a given graph algorithm as the worst case, over all G^{true}, of the expected number of steps taken by the algorithm applied to a slightly randomly perturbed ("smoothed") graph derived from G^{true}.

1.1 The framework

We model a network as an *edge-weighted* graph $G = (\mathcal{V}, \mathcal{E}, \mathbf{w})$. Having in mind social networks, the edge-weights $\mathbf{w} = (w_e : e \in \mathcal{E})$ are regarded as "strength of association" between the entities modeled as vertices; note this is the opposite convention from regarding weights as "distance" or cost, which is implicit in concepts such as *minimum spanning tree*. It is plausible that strongly associated edges are easier to observe than weakly associated edges. To model this, we imagine that what is observable is some kind of pairwise interaction between entities, and that interactions across edge e occur at times of a Poisson (rate w_e) process, independently over different edges. (In other words we identify "strength of association" as being "frequency of interaction".) So by time t we have observed a random number $M_e(t)$, with Poisson($w_e t$) distribution, of interactions across e.

In our framework there is an unknown G^{true} with known vertex-set $\mathcal{V}$ but unknown edge-weights $\mathbf{w}$. Note that we can express our observations in two equivalent ways, either as the random multigraph $\mathbf{M}(t)$ with $M_e(t)$ copies of edge e, or as the random weighted graph $G^{\text{obs}}(t)$ in which edge e has weight $t^{-1}M_e(t)$. Although logically equivalent, we shall see that these two representations suggest different questions and techniques. We call $(\mathbf{M}(t), 0 \le t < \infty)$ the *observed multigraph process* and we call $(G^{\text{obs}}(t), 0 \le t < \infty)$ the *observed network process*.

Fig. 1 Equivalent representations of the observed process.



$\mathbf{M}(t)$ $G^{\text{obs}}(t)$

1.2 Estimating functionals

Repeating our initial project description, let us regard the network G^{true} as unknown, and suppose we are given a functional Γ on the space $\mathbb{G}$ of networks (finite edge-weighted graphs): how do we use the observed process to estimate $\Gamma(G^{\text{true}})$? Of course $N_e(t)/t$ is the natural frequentist estimator of w_e, and so $G^{\text{obs}}(t)$ is an estimator of G^{true}, and so we could use $\Gamma(G^{\text{obs}}(t))$ as an estimator for $\Gamma(G^{\text{true}})$. We call this the "naive frequentist estimator", using *naive* as a reminder that there is no reason to believe it is optimal, and we will see an example (section 3) where it is clearly not.

Write the total interaction rate of vertex v as

$$w_v = \sum_y w_{vy}.$$

In informal discussions of weighted graphs the relevant distinctions are somewhat different from the familiar *sparse, dense* distinction for unweighted graphs. Write

$$w^* := \max_v w_v, \quad w_* = \min_v w_v.$$

For a sequence of weighted graphs with $|\mathcal{V}| = n \to \infty$ we envisage that weights have been scaled to make

$$w^* = \Theta(1).$$

Then we can distinguish between

- the *diffuse* case where $\lim_n \max_e w_e = 0$
- the *local-compact* case where $\lim_{\varepsilon \downarrow 0} \limsup_n \max_v \sum\{w_{vy} : \ w_{vy} \le \varepsilon\} = 0$.

See section 6.2 for some background. We also envisage

$$w_* = \Omega(1).$$

It is now conceptually useful to consider three time regimes for the observation process.

Short-term: $t = o(1)$**.** In this regime we see no interactions at a typical vertex. The only aspects of the unknown G we can estimate relate to "local" statistics, such as the (edge-weighted analog of – see section 6.3) degree distribution and densities of triangles or other $O(1)$-size subgraphs ("motifs" in the applied literature).

Long term: $t = \Omega(\log n)$**.** This is the observation time typically required for the observed graph to be connected. After this time we will, in the context of local-compact networks, have good estimates of most edge-weights, and so we expect that $\Gamma(G^{\rm obs}(t))$ will be a good estimator for $\Gamma(G)$, for most functionals Γ.

Medium term: $t = \Theta(1)$**.** This is what we regard as the "interesting case" – informally,

> What can we infer about the unknown network when we have observed an average of (say) 24 interactions per vertex?

This article is intended as first steps of analysis in this framework, by indicating what can be done using two different techniques. The most straightforward technique involves using the estimator $\Gamma(G^{\rm obs}(t))$ or variants, and relies on large deviation bounds for Poisson distributions. We give results for a "community size" functional in section 2.1 and for maximum-weight matching in section 3. These require mild assumptions on the interaction rates w_v of $G^{\rm true}$. A second technique exploits a certain monotonicity property of the observed multigraph process, that for certain stopping times T one can show that the variability s.d.$(T)/\mathbb{E}T$ is bounded *uniformly* over all networks. The implies that $\mathbb{E}T$ is a functional of the network that can be estimated by T. This is a kind of "backwards" technique, in that such functionals may not be very natural in themselves, but one can then seek to relate them to more natural ones. This second technique and some simple examples (involving times to observe triangles or spanning trees) were introduced in [3] and are reviewed in section 4. Such results suggest a more detailed formulation of our estimation program, as follows.

> Given a statistic Γ, define a ("universal") stopping rule T and an estimator $\widehat{\Gamma}(G^{\rm obs}(T))$ such that the relative error of the estimator, that is $\widehat{\Gamma}(G^{\rm obs}(T))/\Gamma(G^{\rm true}) - 1$, is small **uniformly** over all networks $G^{\rm true}$.

Subject to this requirement we want T to be small, but inevitably the size of T will depend on $G^{\rm true}$.

The requirement that estimates be uniformly good over all finite networks of all sizes makes this a very challenging program. This article presents only rather limited results, and is intended to suggest possible further research.

A key open problem in this formulation involves connectivity in the medium term regime. We expect that at (large) times $t = O(1)$, the observed $G^{\rm obs}(t)$ will have a (large) giant component, of some size $(1 - \delta)n$. We seek a result which says that, if we observe some quantitative "well-connected" property

within the giant component of $G^{\rm obs}(t)$, then we can infer that G has some similar connectivity property within some large subset of vertices. This seems intuitively very plausible, but also seems difficult to formalize. We give a weak indirect version, involving multicommodity flow, in section 4.1, but we expect there are more natural versions. The logic of such arguments is rather counter-intuitive, as indicated in section 4.2.

In section 5 we discuss first-passage percolation, as a basic model for spread of information on networks, in our framework. Further general discussion is postponed to section 6.

2 Estimators guaranteed by large deviation bounds

Consider a functional of the form

$$\Gamma(G) = \max_{A \in \mathcal{A}} \sum_{e \in A} w_e$$

where $\mathcal{A}$ is a collection of edge-sets A. For such functionals it does seem reasonable to use $\Gamma(G^{\rm obs}(t))$ as an estimator of $\Gamma(G^{\rm true})$, because the individual sums $\sum_{e\in A} M_e(t)$ have Poisson$(t\sum_{e\in A} w_e)$ distribution which is concentrated around its mean. We study two examples of such functionals, in sections 2.1 and 3.

First we record the elementary large deviation bounds for a Poisson(λ) r.v. Poi(λ). Define

$$-\phi(a) = a - 1 - a\log a, \quad 0 < a < \infty$$

so that $\phi(a) > 0$ for $a \neq 1$. Then

$$\lambda^{-1}\log \mathbb{P}(\mathrm{Poi}(\lambda) \le a\lambda) \le -\phi(a), \quad 0 < a < 1 \tag{1}$$

$$\lambda^{-1}\log \mathbb{P}(\mathrm{Poi}(\lambda) \ge a\lambda) \le -\phi(a), \quad 1 < a < \infty. \tag{2}$$

2.1 Weights of communities

A *community* in a network is, conceptually, a subset of vertices which is better connected than a typical subset of the same size. Algorithms for "community detection" have been a major field of study [6,7] but we consider only maximal sizes of communities and disregard computational complexity issues.

For a subset A^* of vertices write A for the set of edges with both end-vertices in A^*. Write

$$\mathbf{w}_m = \max\left\{\sum_{e\in A} w_e \ : \ |A^*| = m\right\}.$$

How can we estimate this in our framework, where $\mathbf{w} = (w_e, e \in \mathcal{E})$ is unknown? Ignoring computational complexity, suppose we can compute the analogous observable quantity

$$W_m(t) = \max\left\{ \sum_{e\in A} N_e(t)/t \ : \ |A^*| = m \right\}.$$

Typically $W_m(t)$ will be larger than $\mathbf{w}_m$, and for fixed m will typically grow to ∞ as $n \to \infty$ (here we envisage the case where all vertices v have interaction rate w_v of order 1). We interpret "community" as a subset A^* of some size $m = m(n)$ for which $m^{-2}\sum_{e\in A} w_e$ (the average interaction rate between community members) is not $o(1)$. In other words, saying that communities of size m exist is saying that $m^{-2}\mathbf{w}_m$ is not $o(1)$.

Consider the case where the size m is order $\log n$. In this range, the straightforward "first moment" calculation below shows that as t grows the estimation error (when using $W_m(t)/m^2$ to estimate $\mathbf{w}_m/m^2$) decreases as $t^{-1/2}$ uniformly over n and weighted graphs.

The calculation. Because there are $\binom{n}{m}$ subsets of size m,

$$\mathbb{P}(W_m(t) \geq \beta m^2) \leq \binom{n}{m} \mathbb{P}(\mathrm{Poi}(\mathbf{w}_m t) \geq \beta m^2 t).$$

So provided $\mathbf{w}_m < \beta m^2$ we can use the large deviation upper bound (2) to write

$$\begin{aligned} \log \mathbb{P}(W_n(t) \geq \beta m^2) &\leq \log \binom{n}{m} - \mathbf{w}_m t \phi(\beta m^2/\mathbf{w}_m) \\ &\leq m \log n - \mathbf{w}_m t \phi(\beta m^2/\mathbf{w}_m) - \log m! \end{aligned}$$

Now set $m = \gamma \log n$ and $\mathbf{w}_m = \alpha m^2$ for $\alpha < \beta$, and then

$$\log \mathbb{P}(W_n(t) \geq \beta m^2) \leq (\gamma - \gamma^2 \alpha t \phi(\beta/\alpha)) \log^2 n - \log m!$$

So if we take $\beta = \beta(\alpha, \gamma, t)$ as the solution of

$$\gamma \alpha t \phi(\beta/\alpha) = 1 \tag{3}$$

then

$$\mathbb{P}(W_m(t) \geq \beta m^2) \leq 1/m!$$

This tells us that (for $m = \gamma \log n$ and outside an event of probability $\to 0$ as $n \to \infty$) the estimation error $m^{-2}(W_m(t) - \mathbf{w}_m)$ is at most $\beta - \alpha$, for $\alpha = \mathbf{w}_m/m^2$ and β defined by (3).

The conceptual point is that the bounds above are uniform over all networks. To express in more informal but more readily interpretable terms, note $\phi(a) \sim (a-1)^2/2$ as $a \downarrow 1$, which implies that

$$\beta - \alpha \sim \sqrt{\tfrac{2\alpha}{\gamma t}} \text{ as } t \to \infty.$$

So the conclusion is that, upon observing the value of $m^{-2}W_m(t)$, we can be confident that $m^{-2}\mathbf{w}_m$ is in a certain interval which is approximately

$$\left[m^{-2}W_m(t) - \sqrt{\tfrac{2m^{-2}W_m(t)}{\gamma t}}, m^{-2}W_m(t)\right]$$

where $\gamma = m/\log n$.

3 Maximum matchings

Take n even and work with the complete graph by assigning weight zero to edges e outside $\mathcal{E}$. A *matching* is a set π of $n/2$ edges such that each vertex is in exactly one edge. The *weight* of the matching is $\text{weight}(\pi, \mathbf{w}) := \sum_{e\in\pi} w_e$. The *maximum-weight* is $\mathit{\Gamma}_1(\mathbf{w}) := \max_\pi \text{weight}(\pi, \mathbf{w})$. Readers familiar with the notion of *minimal* matchings should recall that in our setting, large edge-weights indicate closeness, not distance.

In our framework the weights $\mathbf{w}$ are unknown. Can we estimate $\mathit{\Gamma}_1(\mathbf{w})$ from the observed $G^{\text{obs}}(t)$ at (large) times $t = O(1)$? The "natural" estimator $\mathit{\Gamma}_1(G^{\text{obs}}(t))$ is unsatisfactory for the following reason. As usual, for informal discussion we imagine graphs G^{true} with w_v of order 1. For a local-compact such graph, $\mathit{\Gamma}_1(\mathbf{w})$ will be order $\Theta(n)$. Suppose instead G^{true} is the complete graph with $w_e = 1/n\ \forall e$ (a prototypical *diffuse* graph), for which $\mathit{\Gamma}_1(\mathbf{w}) = 1/2$. Here $G^{\text{obs}}(t)$ is essentially the Erdős-Rényi random graph $\mathcal{G}(n, t/n)$ with edge-weights $1/t$, and by considering matchings on that graph we have $\mathit{\Gamma}_1(G^{\text{obs}}(t)) \sim c(t)n$ for a certain function $c(t)$ [8]. So, even though $\mathit{\Gamma}_1(G^{\text{obs}}(t))$ might be a good estimator of $\mathit{\Gamma}_1(\mathbf{w})$ for a local-compact graph, if we superimpose a local-compact and a diffuse graph then we see that $\mathit{\Gamma}_1(G^{\text{obs}}(t))$ contains a spurious contribution of order n from the diffuse part.

We will circumvent this issue as follows. First, we say that our goal is to estimate $n^{-1}\mathit{\Gamma}_1(\mathbf{w})$, the weight-per-vertex of the maximum-weight matching, up to small *additive* error; this effectively means we will be able to ignore edges of weight $o(1)$. We then avoid the difficulty above by only using edges for which we have observed at least two "interactions". That is, we define

$$\text{weight}_2(\pi, G^{\text{obs}}(t)) := t^{-1}\sum_{e\in\pi} M_e(t)1_{\{M_e(t)\geq 2\}}$$

$$\mathit{\Gamma}_2(G^{\text{obs}}(t)) := \max_\pi \text{weight}_2(\pi, G^{\text{obs}}(t))$$

and our goal is to show

$n^{-1}\left|\mathit{\Gamma}_2(G^{\text{obs}}(t)) - \mathit{\Gamma}_1(\mathbf{w})\right|$ is small for large t, uniformly over $\mathbf{w}$.

The best we can hope for is an $O(t^{-1/2})$ bound: consider the graph with only one edge.

We will give one result under the assumption that G^{true} satisfies

$$w_e \leq 1\ \forall e \in \mathcal{E} \tag{4}$$

which implies $\Gamma_1(\mathbf{w}) \le n/2$, and another result under the stronger assumption

$$w_v \le 1 \ \forall v \in \mathcal{V}. \tag{5}$$

Proposition 1 *Under assumption (4) we have a lower bound*

$$\mathbb{E}\left[n^{-1}(\Gamma_2(G^{\rm obs}(t)) - \Gamma_1(\mathbf{w}))\right]^- \le t^{-1/2} + \tfrac{1}{2t}(1+\log t) \ \forall \mathbf{w} \ \forall t \ge 1. \tag{6}$$

Under assumption (5) we have an upper bound

$$\mathbb{E}\left[n^{-1}(\Gamma_2(G^{\rm obs}(t)) - \Gamma_1(\mathbf{w}))\right]^+ \le \Psi(t) \ \forall \mathbf{w} \tag{7}$$

where $\Psi(t) = O(t^{-1/2}\log t)$ *as* $t \to \infty$.

A complicated explicit expression for $\Psi(t)$ could be extracted from the proof.

In seeking our goal, the main issue is to upper bound $\Gamma_2(G^{\rm obs}(t))$. In doing this the contribution from $o(1)$-weight edges will be bounded using technical Lemma 1, and because there are *only* exponentially many matchings using $\Theta(1)$-weight edges, we can apply standard large deviation bounds to bound the contribution from $\Theta(1)$-weight edges.

3.1 The lower bound

For any fixed matching π, the sum $\sum_{e\in\pi} M_e(t)$ has Poisson$(t \cdot \text{weight}(\pi, \mathbf{w}))$ distribution. Choose and fix some π attaining the maximum in the definition $\Gamma_1(\mathbf{w}) := \max_\pi \text{weight}(\pi, \mathbf{w})$. So

$$\Gamma_2(G^{\rm obs}(t)) \ge \text{weight}_2(\pi, G^{\rm obs}(t))$$

and it suffices to lower bound the right side. Now $\sum_{e\in\pi} M_e(t)$ has Poisson$(t \cdot \text{weight}(\pi, \mathbf{w}) = t \cdot \Gamma_1(\mathbf{w}))$ distribution, which we will be able to lower bound later by (1). First let us consider the difference

$$\sum_{e\in\pi} M_e(t) - t \cdot \text{weight}_2(\pi, G^{\rm obs}(t)) = \sum_{e\in\pi} M_e(t) 1_{\{M_e(t)=1\}}$$

for which

$$\mathbb{E}\left(t^{-1}\sum_{e\in\pi} M_e(t) - \text{weight}_2(\pi, G^{\rm obs}(t))\right) = \sum_{e\in\pi} w_e \exp(-t w_e).$$

We want to upper bound the right side, based on the facts that $0 \le w_e \le 1$ and $\sum_{e\in\pi} w_e = \Gamma_1(\mathbf{w}) \le n/2$. By considering separately the edges e with $w_e \le b$ and the edges with $w_e > b$ we see

$$\sum_{e\in\pi} w_e \exp(-t w_e) \le \tfrac{n}{2} b + \Gamma_1(\mathbf{w}) \exp(-tb).$$

Minimizing the right side over $b \geq 0$ leads to

$$n^{-1} \sum_{e \in \pi} w_e \exp(-t w_e) \leq \tfrac{1}{2t}\ \psi(\tfrac{2t\Gamma_1(\mathbf{w})}{n})$$

where

$$\begin{aligned} \psi(x) &= 1 + \log x, \quad x \geq 1 \\ &= x, \quad 0 < x \leq 1. \end{aligned}$$

To summarize, set

$$D_2 := n^{-1} \left(t^{-1} \sum_{e \in \pi} M_e(t) - \text{weight}_2(\pi, G^{\text{obs}}(t)) \right) \geq 0$$

and we have shown

$$\mathbb{E} D_2 \leq \tfrac{1}{2t}\ \psi(\tfrac{2t\Gamma_1(\mathbf{w})}{n}). \tag{8}$$

As noted above, $\sum_{e \in \pi} M_e(t)$ has Poisson$(t \cdot \Gamma_1(\mathbf{w}))$ distribution, and we are interested in showing the difference from expectation

$$D_1 := n^{-1} \left(t^{-1} \sum_{e \in \pi} M_e(t) - \Gamma_1(\mathbf{w}) \right)$$

(in the negative direction) must be small. So fix $\delta > 0$ and calculate

$$\mathbb{P}(D_1 < -\delta) = \mathbb{P}\left(\sum_{e \in \pi} M_e(t) < t\Gamma_1(\mathbf{w}) - nt\delta \right).$$

Applying (1) with $\lambda = t\Gamma_1(\mathbf{w})$ and $a = 1 - n\delta/\Gamma_1(\mathbf{w})$ gives

$$\mathbb{P}(D_1 < -\delta) \leq \exp\left(-t\Gamma_1(\mathbf{w}) \phi(1 - n\delta/\Gamma_1(\mathbf{w})) \right).$$

Because $-\phi(1-\eta) \leq -\eta^2/2$ we find

$$\mathbb{P}(D_1 < -\delta) \leq \exp\left(-\tfrac{tn^2\delta^2}{2\Gamma_1(\mathbf{w})} \right).$$

Because $n^{-1}\Gamma_1(\mathbf{w}) \leq 1/2$ and $n \geq 2$ we get

$$\mathbb{P}(D_1 < -\delta) \leq \exp(-2t\delta^2).$$

Note that if δ is such that $a < 0$ then the probability is zero, so the bound remains valid. Integrating over δ gives

$$\mathbb{E} \max(0, -D_1) \leq 2^{-3/2} \pi^{1/2} t^{-1/2}. \tag{9}$$

To put this all together,

$$\begin{aligned} D &:= n^{-1} \left(\Gamma_2(G^{\text{obs}}(t)) - \Gamma_1(\mathbf{w}) \right) \\ &\geq n^{-1} \left(\text{weight}_2(\pi, G^{\text{obs}}(t)) - \Gamma_1(\mathbf{w}) \right) = D_1 - D_2 \end{aligned}$$

and so

$$\begin{aligned}\mathbb{E}\max(0,-D) &\le \mathbb{E}\max(0,D_2-D_1)\\ &\le \mathbb{E}D_2+\mathbb{E}\max(0,-D_1)\\ &\le 2^{-3/2}\pi^{1/2}t^{-1/2}+\tfrac{1}{2t}\ \psi(t) \qquad (10)\end{aligned}$$

using (8,9) and using again the inequality $\Gamma_1(\mathbf{w}) \le n/2$. This implies the weaker lower bound stated at (6).

3.2 The upper bound

For any fixed matching π, the sum $\sum_{e\in\pi} M_e(t)$ has Poisson$(t\cdot \text{weight}(\pi,\mathbf{w}))$ distribution, and $\text{weight}(\pi,\mathbf{w}) \le \Gamma_1(\mathbf{w})$, so by the large deviation upper bound (2) with $\lambda = t\Gamma_1(\mathbf{w})$ we have

$$\frac{1}{t\Gamma_1(\mathbf{w})}\log\mathbb{P}\left(\sum_{e\in\pi} M_e(t) \ge nt(\tfrac{\Gamma_1(\mathbf{w})}{n}+a)\right) \le -\phi\left(1+\tfrac{an}{\Gamma_1(\mathbf{w})}\right), \quad a>0.$$

We can rewrite this inequality as

$$n^{-1}\log\mathbb{P}\left(\ n^{-1}\sum_{e\in\pi} M_e(t)/t \ge \tfrac{\Gamma_1(\mathbf{w})}{n}+a\right) \le -tn^{-1}\Gamma_1(\mathbf{w})\phi\left(1+\tfrac{an}{\Gamma_1(\mathbf{w})}\right), \quad a>0. \qquad (11)$$

For integer $k \ge 2$ write Π_k for the set of partial matchings π that use only edges e with $w_e > 1/k$ and are maximal subject to that constraint. We can bound the cardinality of that set crudely as $|\Pi_k| \le k^n$. For any matching π, the subset of edges with $w_e > 1/k$ form part of a partial matching in Π_k, and it follows from (11) and the bound $|\Pi_k| \le k^n$ that

$$\begin{aligned}n^{-1}\log\mathbb{P}&\left(\exists\pi\in\Pi_k\ :\ n^{-1}\sum_{e\in\pi, w_e>1/k} M_e(t)/t \ge \tfrac{\Gamma_1(\mathbf{w})}{n}+a\right) \qquad (12)\\ &\le -tn^{-1}\Gamma_1(\mathbf{w})\phi(1+\tfrac{an}{\Gamma_1(\mathbf{w})})+\log k.\end{aligned}$$

To study the contribution from low-weight edges, write

$$\Delta_k(\pi) := \sum_{e\in\pi, w_e\le 1/k} M_e(t)1_{\{M_e(t)\ge 2\}}.$$

Because a matching uses only one edge at a vertex, we can bound this in the form

$$\max_\pi \Delta_k(\pi) \le \tfrac{1}{2}\sum_v M_v^* 1_{\{M_v^*\ge 2\}}; \quad M_v^* = \max\{M_{vy}(t):\ w_{vy}\le 1/k\}. \qquad (13)$$

We will use the following lemma.

Lemma 1 *Let $(N_i, i \geq 1)$ be independent Poisson(λ_i), and write $N^* = \max_i N_i$. Suppose $s := \sum_i \lambda_i \geq 1$ and choose $\lambda^* \geq 1$ such that $\max_i \lambda_i \leq \lambda^* \leq s$. Then*

$$\mathbb{E}N^* 1_{\{N^* \geq 2\}} \leq C\lambda^* \ (1 + \log(s/\lambda^*)) \tag{14}$$

for some numerical constant C.

We outline a proof below using standard methods; the extensive classical theory of extremes [10] focuses on asymptotics in the i.i.d. setting, but it is hard to locate results like Lemma 1.

Because $M_{vy}(t)$ has Poisson(tw_{vy}) distribution, and $\sum_{y:w_{vy} \leq 1/k} w_{vy} \leq 1$ by assumption (5), we can apply Lemma 1 with $s = t$ and $\lambda^* = t/k$, and (14) shows

$$\mathbb{E}M_v^* 1_{\{M_v^* \geq 2\}} \leq Ctk^{-1}(1 + \log k), \quad k \leq t.$$

Applying (13) gives

$$\tfrac{1}{n}\mathbb{E}[\max_\pi \Delta_k(\pi)] \leq \tfrac{1}{2}Ctk^{-1}(1 + \log k), \quad k \leq t. \tag{15}$$

Recall that our goal is to get an upper bound on

$$D := n^{-1}\left(\Gamma_2(G^{\text{obs}}(t)) - \Gamma_1(\mathbf{w})\right).$$

Write B for the event in (12). On the complement B^c we have

$$n^{-1}\Gamma_2(G^{\text{obs}}(t)) \leq n^{-1}\Gamma_1(\mathbf{w}) + a + n^{-1}t^{-1}\max_\pi \Delta_k(\pi).$$

That is,

$$D \leq a + n^{-1}t^{-1}\max_\pi \Delta_k(\pi).$$

Writing F for the event $\{n^{-1}t^{-1}\max_\pi \Delta_k(\pi) > a\}$ we have

$$D \leq 2a \text{ on } B^c \cap F^c$$

and from Markov's inequality and (15)

$$\mathbb{P}(F) \leq Ck^{-1}(1 + \log k)/a, \quad k \leq t.$$

Recall (12) gave a bound on $\mathbb{P}(B)$. Combining these bounds,

$$\mathbb{P}(D > 2a) \leq \exp\left[n(-tn^{-1}\Gamma_1(\mathbf{w})\phi(1 + \tfrac{an}{\Gamma_1(\mathbf{w})}) + \log k)\right] + Ck^{-1}(1+\log k)/a, \quad k \leq t. \tag{16}$$

We want to optimize over choice of k.

So far we have been precise with the bounds, but for ease of exposition let us continue the calculations considering only the leading terms. In particular, treat the asymptotic relation $\phi(1 + \delta) \sim \delta^2/2$ as exact for small $\delta > 0$. This makes the term

$$\Gamma_1(\mathbf{w})\phi(1 + \tfrac{an}{\Gamma_1(\mathbf{w})}) = \tfrac{a^2 n}{2}\tfrac{n}{\Gamma_1(\mathbf{w})} \geq a^2 n$$

because $\Gamma_1(\mathbf{w}) \le 1/2$. So

$$\mathbb{P}(D > 2a) \le k^n \exp(-nta^2) + Ck^{-1}(1 + \log k)/a, \quad k \le t. \tag{17}$$

Note this bound does not depend on $\mathbf{w}$. Integrate over a to get

$$\int_{a_0}^1 \mathbb{P}(D > 2a)da \le k^n \tfrac{1}{2nta_0} \ \exp(-nta_0^2) + Ck^{-1}(1 + \log k)\log(1/a_0), \quad k \le t. \tag{18}$$

Now set $k = t$ and $a_0 = t^{-1/2}\log t$, for large t. The bound in (18) becomes

$$\frac{\exp(-n(\log^2 t - \log t))}{2nt^{1/2}\log t} + \frac{C\log^2 t}{t}.$$

This is bounded, uniformly in n, by a function which is $o(t^{-1/2})$ as $t \to \infty$. One can check that this conclusion

$$\int_{a_0}^1 \mathbb{P}(D > 2a)da = o(t^{-1/2}) \text{ as } t \to \infty, \text{ uniformly in } n$$

remains true under the asymptotics $\phi(1 + \delta) \sim \delta^2/2$.

Finally, write

$$\mathbb{E}D^+ \le 2a_0 + 2\int_{a_0}^1 \mathbb{P}(D > 2a)da + \int_2^\infty \mathbb{P}(D \ge a)da.$$

To handle the last term, note $D \le n^{-1}\Gamma_2(G^{\text{obs}}(t))$ and use the crude bound

$$\Gamma_2(G^{\text{obs}}(t)) \le t^{-1}\sum_e M_e(t).$$

The sum has Poisson$(t\sum_e w_e)$ distribution, so by (5)

$$D \text{ is stochastically smaller than } \tfrac{1}{nt}\text{Poi}(nt/2)$$

and the elementary large deviation upper bound (2) for Poisson shows that $\int_2^\infty \mathbb{P}(D \ge a)da \to 0$ exponentially fast in nt. We conclude that $\mathbb{E}D^+$ is indeed $O(t^{-1/2}\log t)$ as $t \to \infty$, uniformly in n.

Proof of Lemma 1. Note first that we can represent the N_i as the counts of a rate-1 Poisson point process on $[0, s]$ in successive intervals of lengths λ_i. But consider instead the collection of $k = \lceil s/\lambda^* \rceil$ successive intervals of length λ^*. Each interval in the first collection is contained within the union of two successive intervals of the second collection. So the proof of (14) reduces to the proof of the following special case: there exists a constant C such that, if $(N_i, 1 \le i \le k)$ are i.i.d. Poisson(λ^*) with $\lambda^* \ge 1$, then

$$\mathbb{E}N^* 1_{\{N^* \ge 2\}} \le C\lambda^* \ (1 + \log(k)).$$

But in fact this bound holds for $\mathbb{E}N^*$, as follows. First, it is easy to show there exists a constant $B < \infty$ such that

$$\frac{\mathbb{P}(\mathrm{Poi}(\lambda^*) \geq i+1)}{\mathbb{P}(\mathrm{Poi}(\lambda^*) \geq i)} \leq \tfrac{1}{2}, \quad \lambda^* \geq 1, i \geq B\lambda^*. \tag{19}$$

Now write

$$\begin{aligned}
\mathbb{E}N^* &= \sum_{i \geq 1} \mathbb{P}(N^* \geq i) \\
&\leq \sum_{i \geq 1} \min(1, k\mathbb{P}(\mathrm{Poi}(\lambda^*) \geq i)) \\
&\leq 1 + \max(B\lambda^*,\ \min\{i : \mathbb{P}(\mathrm{Poi}(\lambda^*) \geq i) \leq 1/k\}) \text{ by (19).}
\end{aligned}$$

Now it is enough to show there exists $C^* < \infty$ such that

$$\mathbb{P}(\mathrm{Poi}(\lambda^*) \geq i) \leq 1/k, \quad \lambda^* \geq 1, i \geq C^*\lambda^*\ (1 + \log(k)$$

and this follows easily from the large deviation upper bound (2).

4 Concentration inequalities for the observed multigraph process

Write $\mathbf{m} = (m_e, e \in \mathcal{E})$ for a multigraph on a given vertex-set $\mathcal{V}$; here $m_e \geq 0$ is the number of copies of each edge e linking vertices of $\mathcal{V}$. The observed multigraph process $(\mathbf{M}(t), 0 \leq t < \infty) = (M_e(t), e \in \mathcal{E}, 0 \leq t < \infty)$ in section 1.1 is a continuous-time Markov chain, whose state space is the set $\mathbb{M}$ of all multigraphs over $\mathcal{V}$, and whose transition rates are

$$\mathbf{m} \to \mathbf{m} \cup \{e\}, \quad \text{rate } w_e$$

where $\mathbf{m} \cup \{e\}$ denotes appending another copy of e to $\mathbf{m}$. As a Markov chain we can start $(\mathbf{M}(t)$ at any state, so let us call the observed process in our framework, which starts from the empty set $\emptyset$, the *standard* process.

Consider a stopping time of the following form.

$$T = T_{\mathcal{A}} = \inf\{t :\ \mathbf{M}(t) \in \mathcal{A}\} \tag{20}$$

where $\mathcal{A} \subset \mathbb{M}$ is a set of multigraphs with the "increasing" property

$$\text{if } \mathbf{m} \in \mathcal{A} \text{ then } \mathbf{m} \cup \{e\} \in \mathcal{A}\ \forall e. \tag{21}$$

For the chain started at an arbitrary state $\mathbf{m}$, write the expectation as

$$h(\mathbf{m}) := \mathbb{E}_{\mathbf{m}} T_{\mathcal{A}}.$$

The monotonicity property (21) implies that for any transition $\mathbf{m} \to \mathbf{m} \cup \{e\}$ we have $h(\mathbf{m} \cup \{e\}) \leq h(\mathbf{m})$. In this setting it is not difficult to establish the following simple concentration bound.

Proposition 2 ([3]) *For the standard chain, for a stopping time T of form (20,21),*

$$\frac{\text{var } T}{\mathbb{E}T} \le \max_{\mathbf{m},e}\{h(\mathbf{m}) - h(\mathbf{m} \cup \{e\}) \ : w_e > 0\}.$$

Here are two applications where the bound in Proposition 2 can be estimated nicely. Consider

$$T_k^{\text{tria}} = \inf\{t : \ \mathbf{M}(t) \text{ contains } k \text{ edge-disjoint triangles}\}.$$

$$T_k^{\text{span}} = \inf\{t : \ \mathbf{M}(t) \text{ contains } k \text{ edge-disjoint spanning trees}\}.$$

Proposition 3 ([3])

$$\frac{\text{s.d.}(T_k^{\text{tria}})}{\mathbb{E}T_k^{\text{tria}}} \le \left(\frac{e}{e-1}\right)^{1/2} k^{-1/6}, \ k \ge 1.$$

$$\frac{\text{s.d.}(T_k^{\text{span}})}{\mathbb{E}T_k^{\text{span}}} \le k^{-1/2}, \ k \ge 1.$$

So here the bounds are independent of $\mathbf{w}$, meaning that we can estimate the statistics $\mathbb{E}T_k$ without assumptions on $\mathbf{w}$ by simply observing T_k itself.

So the "backwards" approach is to seek some T in the observed multigraph process which is concentrated around its mean, independent of $\mathbf{w}$, which therefore provides a "uniform over $\mathbf{w}$" estimator of the functional $\Gamma(\mathbf{w})$ defined by the expectation.

The calculations for the bounds in Proposition 3 exploit some special structure of spanning trees and of triangles (though the latter can extended to analogs for any finite "motif"). However these are not very natural functionals. It is an open question whether analogous bounds hold for other "contains k copies" types of structure. This seems plausible in many cases, but we indicate one case where it does not seem to work easily in section 5.1.

One can weaken the condition that the *maximum* $\max_{\mathbf{m},e}\{h(\mathbf{m}) - h(\mathbf{m} \cup \{e\}) \ : w_e > 0\}$ be bounded to a condition that for "most" possible transitions this is bounded. See applications in [3] to a first-passage percolation question, and in [1] to the appearance of the incipient giant component in inhomogeneous bond percolation, though these problems are outside the framework of this article.

As an alternative to Proposition 2, in the setting of (20,21) we clearly have the submultiplicative property

$$\mathbb{P}(T > t_1 + t_2) \le \mathbb{P}(T > t_1)\ \mathbb{P}(T > t_2), \quad t_1, t_2 > 0. \tag{22}$$

It is well known that (22) implies a right tail bound

$$\sup\{\mathbb{P}(\tfrac{T}{\mathbb{E}T} > t) : \ T \text{ submultiplicative }) \text{ decreases exponentially as } t \to \infty.$$

Note also there is a left tail bound. Because $\mathbb{P}(T > kt_1) \le (\mathbb{P}(T > t_1))^k$ we have $\mathbb{E}T \le t_1/\mathbb{P}(T \le t_1)$, that is $\mathbb{P}(T \le t_1) \le t_1/\mathbb{E}T$, which can be rewritten as

$$\mathbb{P}(T \le a\mathbb{E}T) \le a, \quad 0 < a \le 1.$$

In the language of confidence intervals, this says that (given (22)) after observing the value of T

$$\text{we can be } (1-a)\text{-confident that } \mathbb{E}T \le T/a. \tag{23}$$

Note this is not the "confidence" version of Markov's inequality, which is

$$\text{we can be } (1-a)\text{-confident that } \mathbb{E}T \ge aT.$$

4.1 Connectivity via multicommodity flow

As mentioned in the introduction, a key open problem is to prove a result of the following type. We expect that at (large) times $t = O(1)$, the observed $G^{\text{obs}}(t)$ will have a (large) giant component, of some size $(1-\delta)n = (1-\delta)|\mathcal{V}|$, but will not be completely connected. We seek a result which says that, if we observe some quantitative "well-connected" property within the giant component of $G^{\text{obs}}(t)$, we can infer that G has some similar connectivity property within some large subset of vertices. A common way to quantify connectivity is via the spectral gap of the graph Laplacian. Proving anything like this involving the (restricted) spectral gap – in our context of placing minimal assumptions on $\mathbf{w}$ – seems very difficult. But to show this program is not hopeless, let us give a very weak result in this format, which is easy to prove. Instead of spectral gap, we measure connectivity in terms of the existence of *flows* whose magnitude is bounded relative to edge-weights. Because we are envisaging a context where $G^{\text{obs}}(t)$ is not connected but has a large component containing *most* vertices, we cannot construct flows between all vertex-pairs, but we can consider flows between *most* vertex-pairs.

A *path* from vertex x to vertex y can be regarded as a set of directed edges; a *flow* $\phi_{xy} = (\phi_{xy}(e), e \in \mathcal{E})$ of volume ν is a function that can be represented as

$$\phi_{xy}(e) = \nu\ \mathbb{P}(e \in \gamma_{xy})$$

for some random path γ_{xy} from x to y. Write $|\phi_{xy}|$ for the volume of a flow. A *multicommodity flow* Φ is a collection of flows $(\phi_{x,y},\ (x,y) \in \mathcal{V} \times \mathcal{V})$, maybe of volume zero. Write

$$\Phi[e] = \sum_{(x,y)} \phi_{xy}(e)$$

for the total flow across edge e.

Fix a parameter $\alpha > 0$ and define a functional $\Gamma_\alpha(\mathbf{w})$ on networks as follows. Consider a multicommodity flow Φ constrained by

$$\text{the volume } |\phi_{xy}| \text{ is at most } n^{-2}, \text{ each } (x,y) \in \mathcal{V} \times \mathcal{V} \tag{24}$$

$$\Phi[e] \le \alpha w_e\ \forall e. \tag{25}$$

Then define $\Gamma_\alpha(\mathbf{w})$ as the maximum total flow subject to these constraints:

$$\Gamma_\alpha(\mathbf{w}) := \max_{\Phi \text{ satisfies (24,25)}} \sum_{(x,y)\in\mathcal{V}\times\mathcal{V}} |\phi_{xy}|.$$

Note that $\Gamma_\alpha(\mathbf{w}) \le 1$. For a connected network, the smallest α for which $\Gamma_\alpha(\mathbf{w}) = 1$ is a parameter that can be used to lower bound the spectral gap: this is the well-known *canonical path* or *Poincaré* method [5].

Let us say a network has the (α, δ)-property if $\Gamma_\alpha(\mathbf{w}) \ge 1-\delta$. Knowing this property holds for small δ is an indirect and somewhat weak quantification of the notion that the network has a large well-connected component. Decreasing α or δ makes the property stronger.

In our program, we want to justify an inference of the form: if the observed network has the (α, δ)-property, then we can be confident that the unknown true network has the (α^*, δ^*)-property for some specified (α^*, δ^*).

Regarding the observed multigraph process $(\mathbf{M}(t), 0 \le t < \infty)$ as networks with integer edge-weights, define

$$T = T_{\alpha,\delta} = \inf\{t :\ \Gamma_\alpha(\mathbf{M}(t)) \ge 1 - \delta\}.$$

So for each realization of the observed process, $\mathbf{M}(T)$ permits a flow with total volume $\ge 1 - \delta$ which satisfies (24) and the analog of (25), that is

$$\Phi[e] \le \alpha M_e(T)\ \forall e.$$

Taking expectation over realizations gives a flow Φ with total volume $\ge 1 - \delta$ which satisfies (24) and

$$\Phi[e] \le \alpha \mathbb{E} M_e(T) = \alpha w_e \mathbb{E}T\ \forall e$$

the equality holding by Wald's identity for the Poisson process. That is,

$$G^{\text{true}} \text{ has the } (\alpha\mathbb{E}T, \delta) \text{ property.}$$

Now $\mathbb{E}T$ depends on G^{true}, but we are in the setting of (20,21) and so we can use the "confidence" statement (23). After observing T,

$$\text{we can be } (1-a)\text{-confident that } G^{\text{true}} \text{ has the } (\alpha T/a, \delta)\text{-property.} \quad (26)$$

We conjecture that Proposition 2 or variants can be used to establish a small bound on $\frac{\text{s.d.}(T)}{\mathbb{E}T}$, which would lead to an improvement on (26).

4.2 On the logic of inference

The logic of (frequentist) statistical inference is often found to be counter-intuitive, so may be worth spelling out in our context. Suppose P is some "desirable" property of a network. If we wish to justify an inference procedure of the format

> *Inference:* if G^{obs} has property P then we are $\geq 95\%$ confident that G^{true} has property P^*

then we need to prove a theorem of the format

> *Theorem:* if G^{true} does not have property P^* then with $\geq 95\%$ probability G^{obs} does not have property P.

Usually with random graph models we are interested in establishing some "desirable" property; paradoxically in our framework we need to show G^{obs} has "less desirable" properties than G^{true}. In particular, in questions about connectivity, the issue is **not** to show that G^{obs} has good connectivity properties (which is typically false).

5 First passage percolation

Many aspects of network science involve some notion of "spread of information", so let us consider a mathematically fundamental model. Consider a network $G = (\mathcal{V}, \mathcal{E}, \mathbf{w})$ with two distinguished vertices v^*, v^{**}. Create independent random variables $(\xi_e, e \in \mathcal{E})$ with Exponential(w_e) distributions, and view ξ_e as the "traversal time" of edge e. Let $X(G)$ be the (random) first passage percolation (FPP) time from v^* to v^{**}, that is the minimum value of $\sum_{e\in\pi} \xi_e$ over all paths π from v^* to v^{**}. We can study the functional

$$\Gamma(G) = \mathbb{E}X(G).$$

How well can we estimate this from the observed process? The following easy result says that $X(G^{\text{obs}}(t))$ is stochastically larger than $X(G^{\text{true}})$.

Lemma 2

$$\mathbb{P}(X(G^{\text{obs}}(t)) \geq x) \geq \mathbb{P}(X(G) \geq x),\ 0 < x < \infty.$$

Before giving the proof let us observe that the interpretation of Lemma 2 is rather subtle for several reasons. First, for any fixed t we have $\mathbb{P}(X(G^{\text{obs}}(t)) = \infty) > 0$ because v^* and v^{**} might not be in the same connected component of $G^{\text{obs}}(t)$. So any plausible estimation procedure would need to continue until some stopping time at which they are in the same component. Unfortunately Lemma 2 apparently does not extend in any simple way to stopping times. Moreover Lemma 2 refers to the *unconditional* distribution of $X(G^{\text{obs}}(t))$, whereas what we can observe at time t is the *conditional* distribution given the realization of $G^{\text{obs}}(t)$.

Proof of Lemma 2. The unconditional distribution of $X(G^{\rm obs}(t))$ is the distribution of the FPP time for which the edge-traversal times $\xi^*_e(t)$ are independent with distributions defined by:

the conditional distribution of $\xi^*_e(t)$ given $M_e(t)$ is Exponential$(M_e(t)/t)$.
So it is enough to show that $\xi^*_e(t)$ stochastically dominates the Exponential(w_e) distribution of ξ_e. But

$$\begin{aligned}\mathbb{P}(\xi^*_e(t) \geq x) &= \mathbb{E}\mathbb{P}(\xi^*_e(t) \geq x | N_e(t)) \\ &= \mathbb{E}\exp(-xN_e(t)/t) \\ &\geq \exp(-x\mathbb{E}(N_e(t)/t)) = \exp(-xw_e)\end{aligned}$$

using Jensen's inequality.

5.1 A general conjecture fails

It is clear that we can always use the observation process itself to simulate the FPP process; that is, there is a stopping time T for the observation process which has itself the distribution of $X(G)$. On the other hand for special classes of network we can estimate the mean $\Gamma(G) = \mathbb{E}X(G)$ much more quickly. For instance in a linear graph G on m edges where we know each edge-weight is $\Theta(1)$ we have $\Gamma(G) = \Theta(m)$ but we can estimate it in time $\Theta(\log m)$ by estimating the individual edge-weights. So it is natural to hope that there exist estimation schemes which

$$\text{on every network } G \text{ require at most } O(\Gamma(G)) \text{ observation time} \tag{27}$$

but which for some class of "nice" networks require substantially less observation time. For instance, by analogy with the examples in Proposition 3 one might hope to require only observation time

$$T_k = \inf\{t : \mathbf{M}(t) \text{ contains } k \text{ edge-disjoint paths from } v^* \text{ to } v^{**}\} \tag{28}$$

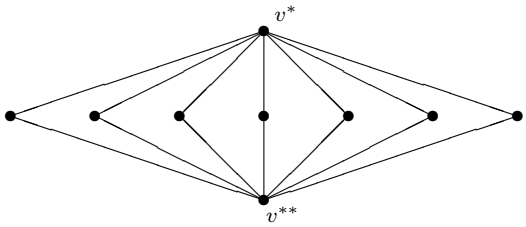
for fixed large k. But this hope is doomed. The argument below, though not completely rigorous, convinces us that

(*) for any estimator satisfying (27), the observation time required must be $\Theta(\Gamma(G))$ (rather than $o(\Gamma(G))$) for *every* G.

However we conjecture that, under mild assumptions on $G^{\rm true}$, one can indeed estimate $\Gamma(G^{\rm true})$ after observation time T_k at (28), analogous to Proposition 3.

Argument. Consider the network $G^{(1)}_n$ as in Figure 2 with n two-edge routes from v^* to v^{**}, and with edge weights $n^{-1/2}$.
Here it is straightforward to see that both "observation time $T^{(1)}_n$ needed" and "actual FPP time $\Gamma(G^{(1)}_n)$" are both $\Theta(1)$. Now suppose we have an estimator satisfying (27). The basis for our argument is the fact that the estimation procedure has to decide whether to stop at time t (and announce an estimate)

Fig. 2 The network $G_n^{(1)}$



or to continue; and it seems intuitively clear that this decision, based on $\mathbf{M}(t)$, can in fact use only the subset $\mathbf{M}^*(t)$ of edges that are in paths in $\mathbf{M}(t)$ from v^* to v^{**}. Because the algorithm cannot make assumptions abut unobserved edges.

So suppose there are networks $\widetilde{G}_n$ for which the estimator needs only observation time $\widetilde{T}_n \ll \Gamma(\widetilde{G}_n)$. We can scale edge-weights so that $\widetilde{T}_n$ is $o(1)$ and $\Gamma(\widetilde{G}_n)$ is $\Omega(1)$. Now define G_n as the superposition of $G_n^{(1)}$ and $\widetilde{G}_n$ – that is, take the union of edges, with the common distinguished vertices (v^*, v^{**}). At time $\widetilde{T}_n$ the estimator will see (with probability $1 - o(1)$) the same set $\mathbf{M}^*(\cdot)$ whether the true network is G_n or $\widetilde{G}_n$. Given it announces a good (that is, $\Omega(1)$) estimate of $\Gamma(\widetilde{G}_n)$ it must announce the same estimate for $\Gamma(G_n)$. But this is incorrect because the availability of paths in $G_n^{(1)}$ means that $\Gamma(G_n)$ is in fact $\Theta(1)$.

6 Final remarks

6.1 Other formulations of imperfectly observed networks

Broad topics around "imperfectly-observed networks" have been studied from many different viewpoints, mostly in the setting of unweighted graphs, and an overview can be gleaned from the talks at the workshop [14]. Here we just mention two such viewpoints. The first is the idea of sampling a few vertices in a large network and looking at their neighborhood structure, which enables one to get estimates of statistics for local structure – see [15] for a recent account. The second is to assume only the possibility of unobserved edges. This is a field called *link prediction*; the 2011 survey [13] cites 166 papers and has been cited 923 times. In this literature, the goal is to define an algorithm that takes the observed edges as input, and outputs an ordering $e_1, e_2, \ldots$ of all the other possible edges, intended as decreasing order of assessed "likelihood" of the edge being present. This is done by defining, for each possible edge (v_1, v_2), some statistic based on (typically) the local structure of the observed graph near v_1 and v_2, for instance

$$s(v_1, v_2) = \frac{|\mathcal{N}(v_1) \cap \mathcal{N}(v_2)|}{|\mathcal{N}(v_1)| \times |\mathcal{N}(v_2)|}$$

where $\mathcal{N}(v)$ is the set of neighbors of v. Then list edges in decreasing order of $s(v_1, v_2)$. However, there is no probability model involved; different algorithms are compared experimentally by taking a real-world network, randomly deleting a proportion of edges to create a synthetic "observed graph", and comparing the algorithms' effectiveness in predicting the deleted edges.

6.2 Convergence of edge-weighted graphs

Recall from section 1.2 that a sequence $G^{(n)} = (\mathcal{V}^{(n)}, \mathcal{E}^{(n)}, \mathbf{w}^{(n)})$ of edge-weighted graphs such that

$$\max_{v \in \mathcal{V}^{(n)}} w_v^{(n)} \text{ is bounded} \tag{29}$$

can be called

- *diffuse* if $\lim_n \max_e w_e^{(n)} = 0$
- *local-compact* if $\lim_{\varepsilon \downarrow 0} \max_v \sum\{w_{vy}^{(n)} : \ w_{vy}^{(n)} \le \varepsilon\} = 0$.

A simple compactness argument shows that we can decompose $\mathbf{w}^{(n)}$ as the sum of two terms, one corresponding to a diffuse sequence and the other to a local-compact sequence. So informally these represent the two possible types of $n \to \infty$ structure for bounded total interaction rate networks.

There is an intuitively natural notion of *local* convergence of finite rooted graphs to a limit locally finite (but typically infinite) rooted graph. One can build upon that notion to define *local weak convergence* of finite unrooted random graphs to a limit locally finite rooted random graph: this merely means taking a uniform random root and applying the previous notion. In the context of unweighted bounded degree graphs this is now known as *Benjamini-Schramm convergence* [4,12]. In fact the notion of local weak convergence extends to edge-weighted graphs under condition (29) rather than bounded-degree: see [2]. (Because *local* means "within fixed distance" we need to reinterpret our edge-weights w_e as lengths $1/w_e$). Without engaging details, the condition for compactness in this topology is essentially our *local-compact* condition above.

6.3 Degree distribution and diffusivity

Our framework is rather different from the "sampling vertices from a graph which can be explored" literature for unweighted graphs [15]. In that framework one can sample k vertices and see their degrees, thereby getting an estimate of degree distribution which has $O(1/\sqrt{k})$ error independent of the graph size n. In our framework the only aspect we can estimate from $O(1)$ observed edges is the total weight $w = \frac{1}{2}\sum_v w_v = \sum_e w_e$. In an edge-weighted graph, one might use the distribution of $W = w_v$ for uniform random $v \in \mathcal{V}$ to play

the role of degree distribution. Assuming W is $\Theta(1)$ as $n \to \infty$, how long does it take to estimate the distribution of W? We can observe

$$Q(i,t) = \text{ number of vertices with } i \text{ observed edges at time } t$$

and for $t = o(1)$ we have

$$\mathbb{E}Q(i,t) \approx \frac{nt^i \mathbb{E}W^i}{i!}.$$

So in order to estimate $\mathbb{E}W^i$ we need $t = \Omega(n^{-1/i})$, in other words we need to see order $n^{1-1/i}$ edges in total. The upshot is that to estimate the distribution W well we need to see $n^{1-o(1)}$ edges, that is time $t = n^{-o(1)}$.

Somewhat similarly, at a (small) time $t = \Theta(1)$, the mean number of observed repeated edges is approximately $\sum_e w_e^2 t^2/2$, and so the notion above of a *diffuse* network corresponds roughly to this mean number being $o(n)$ rather than $\Theta(n)$.

Acknowledgements A slightly expanded version of this article appears in the Ph.D. thesis [11] of the second author.

References

1. David Aldous. The incipient giant component in bond percolation on general finite weighted graphs. *Electron. Commun. Probab.*, 21:Paper No. 68, 9, 2016.
2. David Aldous and J. Michael Steele. The objective method: probabilistic combinatorial optimization and local weak convergence. In *Probability on discrete structures*, volume 110 of *Encyclopaedia Math. Sci.*, pages 1–72. Springer, Berlin, 2004.
3. David J. Aldous. Weak concentration for first passage percolation times on graphs and general increasing set-valued processes. *ALEA Lat. Am. J. Probab. Math. Stat.*, 13(2):925–940, 2016.
4. Itai Benjamini and Oded Schramm. Recurrence of distributional limits of finite planar graphs. *Electron. J. Probab.*, 6:no. 23, 13 pp. (electronic), 2001.
5. Persi Diaconis and Daniel Stroock. Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Probab.*, 1(1):36–61, 1991.
6. Santo Fortunato. Community detection in graphs. *Phys. Rep.*, 486(3-5):75–174, 2010.
7. Lucas G. S. Jeub, Prakash Balachandran, Mason A. Porter, Peter J. Mucha, and Michael W. Mahoney. Think locally, act locally: Detection of small, medium-sized, and large communities in large networks. *Phys. Rev. E*, 91:012821, Jan 2015.
8. R. M. Karp and M. Sipser. Maximum matching in sparse random graphs. In *Foundations of Computer Science, 1981. SFCS '81. 22nd Annual Symposium on*, pages 364–375, Oct 1981.
9. Michael Krivelevich, Daniel Reichman, and Wojciech Samotij. Smoothed analysis on connected graphs. *SIAM J. Discrete Math.*, 29(3):1654–1669, 2015.
10. M. R. Leadbetter, Georg Lindgren, and Holger Rootzén. *Extremes and related properties of random sequences and processes.* Springer Series in Statistics. Springer-Verlag, New York-Berlin, 1983.
11. Xiang Li. *Inference on Graphs: From Probability Methods to Deep Neural Networks.* PhD thesis, U.C. Berkeley, 2017.
12. László Lovász. *Large networks and graph limits*, volume 60 of *American Mathematical Society Colloquium Publications.* American Mathematical Society, Providence, RI, 2012.
13. Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A*, 390:115—1170, 2011.

14. WIND16. Workshop on incomplete networked data. eliassi.org/WIND16.html. Abstracts for March 2016 workshop.
15. Yaonan Zhang, Eric D. Kolaczyk, and Bruce D. Spencer. Estimating network degree distributions under sampling: an inverse problem, with applications to monitoring social media networks. *Ann. Appl. Stat.*, 9(1):166–199, 2015.