Inferring Degrees from Incomplete Networks and Nonlinear Dynamics

Chunheng Jiang , **Jianxi Gao** and **Malik Magdon-Ismail**

Rensselaer Polytechnic Institute, Troy, NY, USA

{jiangc4,gaoj8}@rpi.edu, magdon@cs.rpi.edu

Abstract

Inferring topological characteristics of complex networks from observed data is critical to understand the dynamical behavior of networked systems, ranging from the Internet and the World Wide Web to biological networks and social networks. Prior studies usually focus on the structure-based estimation to infer network sizes, degree distributions, average degrees, and more. Little effort attempted to estimate the specific degree of each vertex from a sampled induced graph, which prevents us from measuring the lethality of nodes in protein networks and influencers in social networks. The current approaches dramatically fail for a tiny sampled induced graph and require a specific sampling method and a large sample size. These approaches neglect information of the vertex state, representing the dynamical behavior of the networked system, such as the biomass of species or expression of a gene, which is useful for degree estimation. We fill this gap by developing a framework to infer individual vertex degrees with both the sampled topology and vertex states. We combine the mean-field theory with combinatorial optimization to learn vertex degrees. Experimental results on real networks with a variety of dynamics demonstrate that our framework can produce reliable degree estimates and dramatically improve existing link prediction methods by replacing the sampled degrees with our estimated degrees.

1 Introduction

Many networks, although large in sizes, are substantially incomplete [Barabási *et al.*, 2002; Szilagyi *et al.*, 2005; Woodward *et al.*, 2010; Chen *et al.*, 2016] when compared with the real systems. To learn the topological characteristics of a complex network, we often need powerful accessing of the vertices and edges of the whole system. However, the associated cost is prohibitive, especially for large-scale networks, e.g., social networks with millions or even billions of nodes. Therefore, we seek a reliable and cost-effective way to estimate graph characteristics from incomplete (sampled) data [Ribeiro and Towsley, 2012; Eden *et al.*, 2018].

The individual vertex degrees provide a cornerstone to have reliable estimations of many graph statistics, including network size [Bunge and Fitzpatrick, 1993; Stumpf *et al.*, 2008; Katzir *et al.*, 2011; Kurant *et al.*, 2012], average degree [Goldreich and Ron, 2008; Dasgupta *et al.*, 2014] and motif counting [Klusowski and Wu, 2018; Kashtan *et al.*, 2004]. The vertex degree measures the lethality of a node in protein-protein interaction networks [Jeong *et al.*, 2001], which is usually unknown if the network is incomplete. Consequently, estimating individual degree has its practical meaning in understanding the cellular function. Nevertheless, it is a challenging problem. Although many studies are concentrating on graph sampling techniques to infer some macroscopic properties of networks, such as the *degree distribution* or the associated *moments of degree distribution* [Frank, 1980; Stumpf and Wiuf, 2005; Achlioptas *et al.*, 2009; Ribeiro and Towsley, 2012; Zhang *et al.*, 2015], very few work on estimating the *individual vertex degrees* [Ganguly and Kolaczyk, 2017] so far. Ganguly and Kolaczyk [2017] derived a class of estimators from a risk-theoretic perspective and provided analytical analysis for them. These estimators take full advantage of the local topology information in the sampled sub-network to infer the individual degrees of the sampled vertices. However, the wide variety in the network structure can make the attempt in vain if we do not have any knowledge of specific global network characteristics, such as species abundance in ecological networks, or gene expression in biological networks. Such crucial information is available, but was not used in the risk-minimization framework.

In our framework, assuming that we are given a set of coupled nonlinear equations governs dynamics of the network, and the steady-states of all the nodes or a subset of nodes under the dynamics have been observed. We propose to estimate the individual vertex degrees from its steady-state without knowing any topology of the network, and we further improve the accuracy of the prediction if we know a tiny percentage of sampled graph regardless of sampling methods, as illustrated in Figure 1. The motivation is that the steady-state of some process (e.g., an epidemic spreading, gene expression) may be observable even though the full network is not observable. For example, in social networks, we may know who got sick, but we may do not know who interacts with whom; in biological networks, we may know the gene expression of all the genes of a patient, but we may not com-
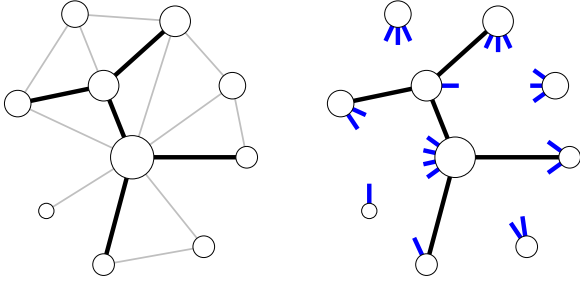
Figure 1: Illustration of our algorithm. (Left), an incomplete network is observed (black edges) and the state values of the dynamics at each vertex is indicated by the size of the vertex. (Right) Our work uses the state values, dynamics plus the incomplete network to infer the degrees at each vertex (blue stubs).

pletely know the wiring diagram of them. It is possible to infer individual vertex degrees because the steady-state is the outcome of a complex interplay between the network's topology and the dynamics. Our framework does not rely on any prior knowledge of the sampling methods or the degree distribution. However, it offers excellent inference with as low as 10% of known edges, and moderate level of noises in the measurements. Another advantage is its ability to recover individual degrees for all vertices with only partial topology, (in)accurate measurements of the state information, though the dynamical system may be misspecified. To demonstrate the impacts of sampling scheme and the measurement noise in steady-states, and how the local topology and the dynamics contribute to the overall accuracy, we conduct a series of empirical experiments on various real networks. Experimental results indicate that our approaches can accurately infer individual vertex degrees and very robust to the sampling scheme. The inference quality can be guaranteed as long as the measurements of the steady-state are not distorted too much.

Problem Setup
Let the implicit full network be $G = (V, E)$. It is only partially observable from a sub-network $G^{(s)} = (V^{(s)}, E^{(s)})$ with $V^{(s)} \subset V$ and $E^{(s)} \subset E$. Let $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_n)$ be the true individual vertex degrees, and $n = |V|$. To be convenient, bold letters hereinafter in mathematical formulas represent vectors.

We consider a general dynamics in which each vertex of G evolves according to a self-driving force and a sum of interaction forces over the neighbors

$$\dot{x}_i = f(x_i) + \sum_{j \in V} A_{ij} g(x_i, x_j), \forall i. \quad (1)$$

Both $f(\cdot)$ and $g(\cdot, \cdot)$ are general and usually nonlinear, and the adjacency matrix $A \geq 0$ modulates the interactions between vertices. The coupled ordinary differential equations (ODEs) which are based on the complete topology of G drive the states of the vertices to an equilibrium $\boldsymbol{x}^*$. Given the complete network, this equilibrium can be approximated efficiently using a mean-field approximation derived in [Gao *et al.*, 2016]

$$\dot{x}_i = f(x_i) + \delta_i g(x_i, x_{\text{eff}}), \forall i. \quad (2)$$

This mean-field approximation interprets the impacts of a neighborhood using an average impact determined by x_{eff}, which is equivalent to a homogeneity assumption. The total interactions of the neighbors of a vertex amounts to multiplying this average interaction by δ_i. To compute the mean-field steady-states $\boldsymbol{z}^*$ in (2), one needs vertex degrees δ_i and the effective network impact, x_{eff}, which can be computed given the complete network. Our contribution is to show how to compute these steady-states given only the incomplete network, thereby establishing a relationship between the observed steady-states and the vertex degrees, from which we back-out the individual degrees. To fill in the details, we first describe the mean-field approximation.

2 The Effective Network Impact

Now consider a vertex i and the interaction term $\sum_j A_{ij} g(x_i, x_j)$ in (1), where A_{ij} is the influence j has on i. Similarly, i influences j with a weight A_{ji}. We define the in-degree $\delta_i^{\text{in}} = \sum_j A_{ij}$ and the out-degree $\delta_i^{\text{out}} = \sum_j A_{ji}$. Assuming $A_{ij} \geq 0$, the interaction can be written as

$$\sum_j A_{ij} g(x_i, x_j) = \delta_i^{\text{in}} \frac{\sum_j A_{ij} g(x_i, x_j)}{\sum_k A_{ik}}. \quad (3)$$

Here, the in-degrees $\boldsymbol{\delta}^{\text{in}}$ captures the idiosyncratic part, and the average $g(\cdot, \cdot)$ captures the network effect. The mean-field approximation is to replace local averaging with global averaging, which approximates the network-impact on a vertex as nearly homogeneous. Specifically, we have

$$\frac{\sum_j A_{ij} g(x_i, x_j)}{\sum_k A_{ik}} \approx \frac{\sum_{ij} A_{ij} g(x_i, x_j)}{\sum_{ik} A_{ik}} = \frac{\mathbf{1}^T A g(x_i, \boldsymbol{x})}{\mathbf{1}^T A \mathbf{1}}, \quad (4)$$

where the vector $g(x_i, \boldsymbol{x})$ has the jth component $g(x_i, x_j)$. We define a linear operator

$$\mathcal{L}_A(\boldsymbol{z}) = \frac{\mathbf{1}^T A}{\mathbf{1}^T A \mathbf{1}} \boldsymbol{z} = \frac{\boldsymbol{\delta}^{\text{out}} \cdot \boldsymbol{z}}{\boldsymbol{\delta}^{\text{out}} \cdot \mathbf{1}} \quad (5)$$

to produce a weighted average of the entries in $\boldsymbol{z}$. Then, in our mean-field approximation,

$$\dot{x}_i = f(x_i) + \delta_i^{\text{in}} \mathcal{L}_A[g(x_i, \boldsymbol{x})]. \quad (6)$$

In the first order linear approximation, we can take the $\mathcal{L}_A$-average inside g. The average of external interactions is approximately the interaction with its average. That is $\mathcal{L}_A[g(x_i, \boldsymbol{x})] \approx g(x_i, \mathcal{L}_A(\boldsymbol{x}))$ and

$$\dot{x}_i = f(x_i) + \delta_i^{\text{in}} g(x_i, \mathcal{L}_A(\boldsymbol{x})), \quad (7)$$

where $\mathcal{L}_A(\boldsymbol{x})$ is a global state. Let $x_{\text{av}} \triangleq \mathcal{L}_A(\boldsymbol{x})$. Applying $\mathcal{L}_A$ to both sides of (7) gives

$$\dot{x}_{\text{av}} = \mathcal{L}_A[f(\boldsymbol{x})] + \mathcal{L}_A[\boldsymbol{\delta}^{\text{in}} g(\boldsymbol{x}, x_{\text{av}})]. \quad (8)$$

According to the extensive discussion and tests in Ref. [Gao *et al.*, 2016], the in-degrees $\boldsymbol{\delta}^{\text{in}}$ and the interaction with the external x_{av} are roughly uncorrelated, so the $\mathcal{L}_A$-average of the product is roughly the product of $\mathcal{L}_A$-averages. Thus, we have $\mathcal{L}_A[\boldsymbol{\delta}^{\text{in}} g(\boldsymbol{x}, x_{\text{av}})] \approx \mathcal{L}_A(\boldsymbol{\delta}^{\text{in}}) \mathcal{L}_A[g(\boldsymbol{x}, x_{\text{av}})]$. Using the

first order linear approximation, we take the $\mathcal{L}_A$-average inside f and g

$$\dot{x}_{\text{av}} = f(\mathcal{L}_A(\boldsymbol{x})) + \mathcal{L}_A(\boldsymbol{\delta}^{\text{in}})g(\mathcal{L}_A(\boldsymbol{x}), x_{\text{av}}). \qquad (9)$$

Now we have

$$\dot{x}_{\text{av}} = f(x_{\text{av}}) + \beta g(x_{\text{av}}, x_{\text{av}}), \qquad (10)$$

where the resilience $\beta = \mathcal{L}_A(\boldsymbol{\delta}^{\text{in}})$. For undirected graphs, $\boldsymbol{\delta}^{\text{in}} = \boldsymbol{\delta}^{\text{out}}$, so $\beta = \sum_i \delta_i^2 / \sum_i \delta_i = \langle \boldsymbol{\delta}^2 \rangle / \langle \boldsymbol{\delta} \rangle$. The steady-state of (10) is the effective network impact, x_{eff}. Plugging it into (7) gives an uncoupled ODE for x_i,

$$\dot{x}_i = f(x_i) + \delta_i^{\text{in}} g(x_i, x_{\text{eff}}). \qquad (11)$$

In the mean-field approximation, $g(x_i, x_j)$ in (1) is replaced by an interaction with a mean-field external world $g(x_i, x_{\text{eff}})$ and the number of neighbors impacting x_i is captured by δ_i^{in}. To approximately obtain the steady-states of the system, one first solves the ODE in (10) to get x_{eff}, and then n uncoupled ODEs at each vertex to get x_i, which only depends on δ_i^{in} if given x_{eff}. The method works well because the mean-field approximations only need to hold *at the steady-state*. Hence, we can recover the steady-state for any vertex (for example the sampled vertices) from accurate estimates δ_i^{in} (δ_i in the undirected case) and β.

3 Our Approaches

Our approaches need the observed steady-states $\boldsymbol{x}^*$ of nodes to estimate the individual node degrees $\boldsymbol{\delta}$, which can be improved if we also know the partial topology of the network. We consider the observed $\boldsymbol{x}^*$ as the ground truths and aim to search an optimal combination of degrees such that the predicted steady-states $\boldsymbol{z}^*$ of the approximated system (2) are close to $\boldsymbol{x}^*$. In this sense, the general techniques underlying our work can be formulated as an optimization problem:

$$\begin{aligned} \min_{\hat{\boldsymbol{\delta}}} \quad & \text{Err}(\boldsymbol{x}^*, \boldsymbol{z}^*) \\ s.t. \quad & \hat{\delta}_i = \delta_i^{(s)} + d_i, \\ & \hat{\beta} = \langle \hat{\boldsymbol{\delta}}^2 \rangle / \langle \hat{\boldsymbol{\delta}} \rangle, \qquad (12) \\ & f(\hat{x}_{\text{eff}}) + \hat{\beta} g(\hat{x}_{\text{eff}}, \hat{x}_{\text{eff}}) = 0, \\ & f(z_i^*) + \sum_j A_{ij}^{(s)} g(z_i^*, x_j^*) + d_i g(z_i^*, \hat{x}_{\text{eff}}) = 0, \end{aligned}$$

where $\delta_i^{(s)}$ is the degree of the sampled vertex i in the incident/induced subgraph, d_i is the estimated missing degree for i (see more details in Section 3.2).

First, we propose the **ZeroTopo** algorithm to estimate the degree of each node without knowing any topology of the network, i.e., $A^{(s)}$ has only zero elements. Then, we design an efficient 2-step procedure – **TopoPlus** and **Round** – to improve the accuracy of degree prediction by introducing the observed topology.

- **ZeroTopo** estimates $\boldsymbol{\delta}$ and x_{eff} with only the observed steady states $\boldsymbol{x}^*$ and with ZERO known topology.

- **TopoPlus** estimates $\boldsymbol{\delta}$ and $\boldsymbol{d}$ with both the partially observed topology and $\boldsymbol{x}^*$. Thus, $\hat{\boldsymbol{\delta}} = \boldsymbol{\delta}^{(s)} + \boldsymbol{d}$, meaning that the estimated degree of a node is equal to the sum of its sampled degree and the missing degree.

Algorithm 1 ZeroTopo

Input: $\boldsymbol{x}^*$
1: Let $t = 0$ and $x_{\text{eff}}^{(0)} = \langle \boldsymbol{x}^* \rangle$
2: Solve $\hat{\delta}_i^{(1)}$ from (2) with $x_{\text{eff}}^{(0)}$
3: **repeat**
4:      Compute $\hat{\beta}^{(t)} = \langle \hat{\boldsymbol{\delta}}^{(t)2} \rangle / \langle \hat{\boldsymbol{\delta}}^{(t)} \rangle$
5:      Solve $x_{\text{eff}}^{(t)}$ from (10) with $\hat{\beta}^{(t)}$
6:      Solve $\hat{\delta}_i^{(t+1)}$ from (2) with $\boldsymbol{x}^*$ and $x_{\text{eff}}^{(t)}$
7: **until** $\hat{\beta}^{(t)}$ and $x_{\text{eff}}^{(t)}$ do not change
Output: $\hat{\boldsymbol{\delta}} = \lfloor \hat{\boldsymbol{\delta}}^{(t)} \rceil$

Algorithm 2 TopoPlus

Input: $\boldsymbol{x}^*, \boldsymbol{\delta}^{(s)}, A^{(s)}$
1: Let $t = 0$ and $\hat{\boldsymbol{\delta}}^{(t)} = \boldsymbol{\delta}^{(s)}$
2: **repeat**
3:      Compute $\hat{\beta}^{(t)} = \langle \hat{\boldsymbol{\delta}}^{(t)2} \rangle / \langle \hat{\boldsymbol{\delta}}^{(t)} \rangle$
4:      Solve $x_{\text{eff}}^{(t)}$ from (10) with $\hat{\beta}^{(t)}$
5:      Solve $d_i^{(t)}$ from (14) with $A^{(s)}, \boldsymbol{x}^*$ and $x_{\text{eff}}^{(t)}$
6:      Update $\hat{\delta}_i^{(t+1)} \leftarrow \delta_i^{(s)} + d_i^{(t)}$
7: **until** $\hat{\beta}^{(t)}$ and $x_{\text{eff}}^{(t)}$ do not change
Output: $\hat{\boldsymbol{\delta}} = \lfloor \hat{\boldsymbol{\delta}}^{(t)} \rceil$, $\boldsymbol{d} = \lfloor \boldsymbol{d}^{(t)} \rceil$ and $\hat{x}_{\text{eff}} = x_{\text{eff}}^{(t)}$

- **Round** further optimizes (12) for *integer* missing degrees $\hat{\boldsymbol{d}}$, and the final estimate becomes $\hat{\boldsymbol{\delta}}^* = \boldsymbol{\delta}^{(s)} + \hat{\boldsymbol{d}}$.

3.1 ZeroTopo

ZeroTopo refers to the estimation of vertex degree based on observed steady-states $\boldsymbol{x}^*$ and the dynamics $f(\cdot)$ and $g(\cdot, \cdot)$, without any known topology information, as shown in Algorithm (1). The average steady-state is set as the initial value of the effective state $x_{\text{eff}} = \langle \boldsymbol{x}^* \rangle$ when **ZeroTopo** warms up (line 1). Based on the mean-field approximation, we can repeatedly estimate the resilience index β from estimated degrees and update the effective state such that the entire system is self-consistent. The algorithm is simple but very powerful to recover all individual degrees simultaneously.

3.2 TopoPlus

In the **ZeroTopo** algorithm, we assume that we have no prior information about the topology of the network. In this section, we aim to enhance the accuracy of the degree prediction when we observe a subgraph $G^{(s)}$. Note that the **TopoPlus** algorithm only requires the topology of the observed subgraph regardless of how one samples it. We incorporate the adjacent matrix $A^{(s)}$ into (2) to identify the missing degree d_i in the discarded local topology. Hence, we get an enhanced mean-field approximation

$$\dot{x}_i = f(x_i) + \sum_j A_{ij}^{(s)} g(x_i, x_j^*) + d_i g(x_i, x_{\text{eff}}), \qquad (13)$$

with the steady-states $\boldsymbol{x}^*$ of the observed vertices, and the effective network impact x_{eff}. Once x_{eff} is found, it is easy to

Algorithm 3 Round

Input: $\boldsymbol{x}^*, \boldsymbol{\delta}^{(s)}, \boldsymbol{d}, \hat{x}_{\text{eff}}$
Output: $\hat{\boldsymbol{\delta}}^* = \boldsymbol{\delta}^{(s)} + \boldsymbol{d}$

1: **repeat**
2:     Let $z_i^*(d_i)$ be a solution to
        $\dot{z}_i = f(z_i) + \sum_j A_{ij}^{(s)} g(z_i, x_j^*) + d_i g(z_i, \hat{x}_{\text{eff}})$
3:     Let $\varepsilon_i(d_i) = \text{Err}\{z_i^*(d_i), x_i^*\}$
4:     Compute $Q_i^+ = \varepsilon_i(d_i + 1) - \varepsilon_i(d_i)$,
        $Q_i^- = \varepsilon_i(d_i - 1) - \varepsilon_i(d_i), \forall i$
        ▷ *isolated nodes are exclusive in calculating Q^-*
5:     Obtain nodes I^+ with $Q_i^+ \leq 0$, I^- with $Q_i^- \leq 0$
6:     Let $I = I^+ \cap I^-$, $I^+ \leftarrow I^+ - I$ and $I^- \leftarrow I^- - I$
7:     **for** each $(i, j) \in (I^+, I^-)$ **do**
8:         $d_i \leftarrow d_i + 1, d_j \leftarrow d_j - 1$
9:     **end for**
10:    **for** each $i \in I$ **do**
11:       **if** $|Q_i^+| > |Q_i^-|$ **then**
12:         $d_i \leftarrow d_i + 1$
13:       **else if** $|Q_i^+| < |Q_i^-|$ **then**
14:         $d_i \leftarrow d_i - 1$
15:       **end if**
16:    **end for**
17:    Let $\hat{\boldsymbol{\delta}}^* = \boldsymbol{\delta}^{(s)} + \boldsymbol{d}$, and calculate $\hat{\beta} = \langle \hat{\boldsymbol{\delta}}^{*2} \rangle / \langle \hat{\boldsymbol{\delta}}^* \rangle$
18:    Solve $\hat{x}_{\text{eff}}$ from (10) with $\hat{\beta}$
19: **until** $\boldsymbol{d}$ does not change

---

derive d_i, that is

$$d_i = -\left[f(x_i^*) + \sum_j A_{ij}^{(s)} g(x_i^*, x_j^*)\right] / g(x_i^*, x_{\text{eff}}). \quad (14)$$

The term d_i is a compensation to $\delta_i^{(s)}$, i.e., $\hat{\delta}_i = \delta_i^{(s)} + d_i$. The idea lies in the mean-field approximation (2) that $\sum_j A_{ij} g(x_i^*, x_j^*) \approx \delta_i g(x_i^*, x_{\text{eff}})$. We do not have *all* connections δ_i associated with vertex i, but we have *some*, $\delta_i^{(s)} \leq \delta_i$. Based on the same approximation (2), we know $\sum_j A_{ij}^{(s)} g(x_i^*, x_j^*) \approx \delta_i^{(s)} g(x_i^*, x_{\text{eff}})$, motivating us to search the missing $d_i = \delta_i - \delta_i^{(s)}$ from (13) and (14). Note that for vertices not in the sampled vertices, $\delta_i^{(s)} = 0, \forall i \notin V^{(s)}$. As a result, the enhanced approximation (13) is equivalent to (2), then computing d_i in (14) is exactly solving $\hat{\delta}_i$ from (2) in **ZeroTopo**. **TopoPlus** is formulated in Algorithm (2).

3.3 Round

Both **ZeroTopo** and **TopoPlus** contain an easily unnoticed but crucial step: **Round**. A rounding algorithm involves the approximation of fractional values to integer ones. In specific, Eq.(14) gives a fractional d_i, and in our setup we consider the networks are unweighted. Therefore, we need the rounding to get an integer approximation of the degree of each node $\hat{\delta}_i = \delta_i^{(s)} + d_i$. It is a complicated and hard problem. In **ZeroTopo** and **TopoPlus** (marked with $\lfloor \cdot \rceil$), we applied the simple plain rounding algorithm because of its ease to use. However, rounding is beyond a purely numerical problem, and we should also consider its afterward impact on the

predicted steady-states $\boldsymbol{z}^*$. Therefore, we design a heuristic combinatorial optimization based rounding scheme (see Algorithm 3) to not only ensure the accuracy in estimating individual degrees, but also minimize the error between the observed steady-states and the simulated steady-state based on the estimated degrees. The idea is to repeatedly pick a node pair (i, j), then adjust their estimated missing degrees d_i and d_j, keeping the total number of links unchanged: increase d_i by 1 and decrease d_j by 1 (line 8, Algorithm 3). The selection rule is that the two steady-states z_i^* and z_j^* solved from Eq. (13) do not yet match the observed states $\boldsymbol{x}^*$ well with the current estimates d_i and d_j. Nevertheless, after the adjustments, the steady-states $\boldsymbol{z}^*$ becomes closer to $\boldsymbol{x}^*$. Though computing seems to be very intensive, each distinct missing degree can involve in solving Eq. (13) (line 2, Algorithm 3) for only once, meanwhile, the function only includes one variable. Thus, it is embarrassingly parallelizable.

Dynamics	Net	n	m	$\langle \boldsymbol{\delta} \rangle$	β
ⓐ Ecological	N6	270	8074	59.81	97.96
	N8	97	972	20.04	32.36
ⓑ Regulatory	ME	2268	5620	4.96	40.78
	TY	662	1062	3.21	13.51
ⓒ Epidemic	EM	1133	5451	9.62	18.69
	FB	4039	88234	43.69	106.57

ⓐ $\dot{x}_i = B + x_i(1 - \frac{x_i}{K})(\frac{x_i}{C} - 1) + \sum_j \frac{x_i x_j A_{ij}}{D + E x_i + H x_j}$

ⓑ $\dot{x}_i = -B x_i^f + \sum_j A_{ij} R \frac{x_j^h}{x_j^h + 1}$

ⓒ $\dot{x}_i = -B x_i + \sum_j A_{ij} R(1 - x_i) x_j$

Table 1: Examples of real systems governed by nonlinear dynamics. We test our algorithms on two real networks for each dynamic. The notations n and m represent the number of vertices and edges, respectively. While $\langle \boldsymbol{\delta} \rangle$ and β denote the average degree and resilience index respectively.

4 Experimental Results

We evaluated the performance of our approaches on six real networks, governed by three dynamical equations (ecological, regulatory, and epidemic) [Gao *et al.*, 2016] (Table 1). For each network, we sample a fraction of the edges and repeat 10,000 times to get statistically significant results. On average, it takes only a few of seconds to converge. We employ the fraction of vertices with a relative error of not greater than 5% as a proxy metric of the estimation accuracy, and the relative error is defined as $\text{Err}[\delta_i, \hat{\delta}_i] = |\log(\hat{\delta}_i / \delta_i)|$. As reported in Table 2, we show the performance of each of our algorithms. Though **ZeroTopo** does not take advantage of the sampled local topology information, the dynamics itself can give a decent estimate of vertex degrees even without the sampled sub-network. The **TopoPlus**, either equipped with plain rounding algorithm or the designed **Round** algorithm, can improve the accuracy of our estimation significantly, indicating that it is vital to utilize some available topology and rounding carefully.

Net	ZeroTopo	TopoPlus			TopoPlus + Round		
	–	10%	20%	30%	10%	20%	30%
N6	58.15	60.17	64.63	69.47	60.23	64.76	**69.65**
N8	57.73	61.97	66.46	71.27	62.74	67.14	**71.80**
ME	86.86	88.02	90.00	91.82	89.13	90.91	**92.59**
TY	86.16	87.13	87.81	88.59	87.76	88.41	**89.14**
EM	59.40	63.20	67.11	71.30	65.77	69.75	**73.77**
FB	59.54	64.05	68.82	73.70	65.36	70.16	**75.03**

Table 2: Estimation accuracy on incident subgraphs with various uniform edge sampling fractions $p \in \{10\%, 20\%, 30\%\}$. ZeroTopo is independent of the fraction, so we identify the number as a dash.

4.1 Sampling Technique

Prior structure-based degree estimation methods [Ganguly and Kolaczyk, 2017] rely on strong assumptions about how to sample the sub-network. Specifically, these methods are designed for the induced graph from uniform sampling. In other words, it does not work for other sampling methods, for example, random walk sampling. Our approaches do not, and they are universal, regardless of sampling methods. To demonstrate this, we compare two very different ways to sample edges: uniform sampling, and random walk sampling [Hu and Lau, 2013]. Figure 2 shows that our approaches are very robust to the sampling scheme, as is expected.
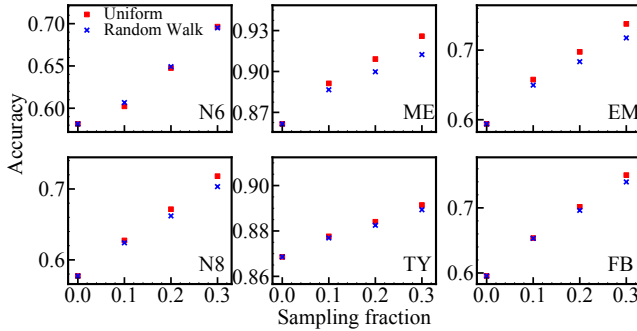
Figure 2: Uniform edge sampling and random walks give similar performance. Sampling fraction $p = 0$ corresponds to our approach that relies on the states only.

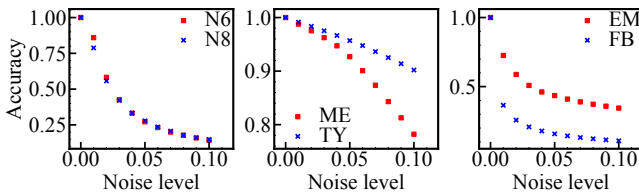## 4.2 Measurement Noise and Model Misspecification

Figure 3: The accuracy decreases as we increase the measurement error σ of the observed steady-states x^* with 10% of sampled edges.

Two fundamental assumptions we made are that the observed steady-states are measured precisely, and the true dy-
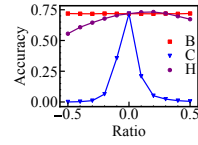
Figure 4: Impact of random perturbation on model parameters.

namics are known. We now show that our approaches can tolerate high level of noise added to the observed steady-states, as well as model misspecification in the form of noise added to the parameters in the dynamical system. We add multiplicative noise $\varepsilon_x \sim N(0, \sigma)$ to the observed states, so we observed the states $x^*(1 + \varepsilon_x) \geq 0$ in simulation. Similarly, the parameters in the dynamical system are perturbed by multiplying by a factor $1 + r$ for $r \in [-0.5, 0.5]$. Figure 3 shows the robustness to observation noise, and Figure 4 shows the impact of errors on three parameters in the ecological system (see Table 2). The parameter B describes the incoming migration rate from neighboring ecosystems, C controls the critical abundance of the Allee effect, and H depicts the positive contribution from one vertex to another. The estimates are very sensitive to the error in the interaction parameters C and H, but is very robust to the error in B.

4.3 Local Topology and Dynamics

We, for the first time, estimate the structure characteristics of graphs using the network dynamics. Then, we compare the performance of our estimators and that of the structure-based estimators [Ganguly and Kolaczyk, 2017]: the univariate risk minimizer (**URM**), the multivariate risk minimizer (**MRM**) and the **Bayes** estimator as the baselines. Note that our estimator is regardless of the sampling methods, and the structure-based methods are only suitable for induced subgraph sampling. To ensure the fairness of the comparison, for each run, (1) we uniformly sample the same 10% of vertices in our approaches and the structure-based estimators; (2) we only utilize the steady-state of the observed nodes in our approaches and the induced graph of the observed nodes in the structure-based estimators. We perform 10,000 runs and show the average accuracies of the baselines and our approaches in Table 3.

We have demonstrated the promising performance of our approaches on the incident subgraphs. As we discussed in Sec. 4, they can perform well no matter how the sub-network are sampled. Therefore, they are expected to accurately estimate vertex degrees on induced subgraphs as well. Moreover, in Algorithm 1, we assume that the states of all the nodes are known. In this comparison, we only know the states of the observed nodes. To bridge this gap, we have to estimate the global β from the induced subgraph. Fortunately, Jiang et. al [2020] offered a methodology to estimate β from incomplete networks, it can be used here. Without the states of all vertices, as shown in Table 3, the performance of the **ZeroTopo** is still outstanding compared with structure-based methods, even with 10% noises in the measurement. It indicates that the local degrees or the neighborhood in the induced subgraph alone are not sufficient to obtain reliable estimates of individual vertex degrees. Vertex states should also

| Net | 10% **ind**uced **sub**graph | | | 0% ind. sub. | 10% ind. sub. | |
	URM	MRM	Bayes	ZeroTopo (σ)	TopoPlus (σ)	+ Round (σ)
N6	11.01	12.88	12.95	48.33 (0%) 13.03 (10%)	50.57 (0%) 13.33 (10%)	50.67 (0%) 13.30 (10%)
N8	8.28	5.30	5.10	47.42 (0%) 12.63 (10%)	47.93 (0%) 12.68 (10%)	48.69 (0%) 12.68 (10%)
ME	0.63	6.27	42.79	86.93 (0%) 72.82 (10%)	88.70 (0%) 73.97 (10%)	89.82 (0%) 75.35 (10%)
TY	0.54	41.74	19.96	87.16 (0%) 84.27 (10%)	87.73 (0%) 84.71 (10%)	88.06 (0%) 85.08 (10%)
EM	3.74	15.00	6.61	60.39 (0%) 28.30 (10%)	63.95 (0%) 29.30 (10%)	66.27 (0%) 30.63 (10%)
FB	8.23	8.81	9.00	59.51 (0%) 9.19 (9%)	64.03 (0%) 9.52 (9%)	65.42 (0%) 9.91 (9%)

Table 3: Estimation accuracy on induced subgraphs. The numbers in parentheses denote the additional measurement noises σ. Note that Bayes relies on the degree distribution, and we use the real one.

be explored. There is an issue in the steady states: the measurement errors. It is ubiquitous in the real-world that the observed states are suffering from measurement noises. No measurement error (i.e., $\sigma = 0$) is an ideal case, and our approaches are significantly better than the baselines. After adding a moderate amount of noise ($\sigma = 10\%$ on N6, N8, ME, TY, and EM; 9% on FB), our approaches can still beat them. The **TopoPlus** and **Round** provide additional improvement because of the extra topology and careful rounding.

4.4 Application: Link Prediction

Estimating degrees itself is an essential and exciting issue in network science. It is also useful for many other applications, such as link prediction, network reconstruction, identification of essential genes, influencer discovery, and many more. Here, we take link prediction as an example to show how our approach can enhance the current link prediction methods. The most intuitive approach to build links is computing the affinity between a node pair, e.g., (a) Adamic-Adar (AA) [Adamic and Adar, 2003] $s_{u,v} = \sum_{w \in \Gamma_u \cap \Gamma_v} 1/\log \delta_w$, where Γ_u is the neighbors of vertex u, and (b) Preferential Attachment (PA) [Barabási and Albert, 1999] $s_{u,v} = \delta_u \delta_v$. Both metrics contain the degree terms. Therefore, we improve these methods by replacing the observed degrees $\boldsymbol{\delta}^{(s)}$ by the estimated degrees $\hat{\boldsymbol{\delta}}$ using our estimators. We follow a simple rule to scale up the size of the common neighbors $|\Gamma_u \cap \Gamma_v|$ by $\alpha_{u,v} = \min\{\hat{\delta}_u/\delta_u^{(s)}, \hat{\delta}_v/\delta_v^{(s)}\}$.

The area under the receiver-operating characteristic curve (AUC) [Hanley and McNeil, 1983] is used to measure the prediction accuracy. AA requires a joint neighborhood, but PA does not. A small sampling fraction gives a very sparse subnetwork such that joint neighborhoods are absent from $G^{(s)}$. As a result, the affinities of many node pairs are distorted. To evaluate the performance, we randomly sample $p =1\%$ edges for an incident subgraph and predict the remaining 99% missing links. As reported in Table 4, the link prediction of AA is almost random. The scale-up factor can not exercise an impact on AA at all since no joint neighbors are observed. For PA, once the individual degrees are accurately revealed, the link prediction can be of high quality as well. It can give a

promising improvement on all six real networks given as low as 1% observed edges. It demonstrates the value of our algorithm in link prediction. Note that our approach can not only improve the current link prediction methods, but it is also essential for many other inferring issues in network science.

| Metrics | Status | Networks (Sampling $p =1\%$ of edges) | | | | | |
		N6	N8	ME	TY	EM	FB
AA	*naive*	50.22	50.06	50.00	50.00	50.02	50.27
	revised	50.22	50.06	50.00	50.00	50.02	50.27
	truth	96.79	95.64	46.60	54.92	86.63	99.47
PA	*naive*	69.07	59.90	64.13	58.00	56.18	69.44
	revised	89.29	87.11	81.18	91.28	82.33	83.78
	truth	89.40	87.42	81.47	92.45	82.59	83.83

Table 4: AUC of AA and PA, for using the observed topological information only (*naive*), the estimated degrees from our algorithm (*revised*) or the ground-truth topological information of the sampled vertices (*truth*). A simple scale up is made for the *revised* AA.

5 Discussion

We addressed a prevalent problem. Consider this scenario. A biologist has measured the gene expressions of all genes and collected some relationships summarized in the adjacency matrix $A^{(s)}$. The biologist also knows how genes interact, the dynamical system (1). The biologist would like to identify the essential genes using the degree centrality and to make some new discoveries. Then he/she tries to use some structure-based methods to estimate the missing degrees and finds the improvement is limited. The biggest issue is that both the gene expression data and the dynamical systems are known, but he/she does not use them. In the current-state-of-the-art, no methods are available to utilize these information. The difficulty is rooted in lacking methods to theoretically solve the n coupled nonlinear ODE equations with incomplete graph. We bridge this gap by using a mean-field approximation, where the original system can be condensed to a single dimensional system with variable x_{eff} and topological parameter β. We provide a framework to estimate the missing degrees accurately, which enables us to measure the essential genes. Moreover, our approaches are robust to sampling methods and work even without knowing any topology.

There are several directions for future work. First, we can go for a reconstruction of the full network topology, not just vertex degrees, which needs to solve the full link-prediction problem. This is ongoing work. Second, the mean-field approximation is the basis of our algorithm. However, the estimation itself relies on the assumption that the network is homogeneous. Therefore, when the steady-states of the neighbors are unknown, we use x_{eff} as an approximation. All nodes are treated without any discrimination, although their neighborhoods may be different from each other. This assumption may not hold in real-world systems, and the estimation accuracy will be negatively affected. Another issue is the dynamical equations. In many scenarios, it is tough to calibrate the real dynamical equations. We can employ advanced machine learning approaches to extract the governing equations from an immense amount of experimental data.

References

[Achlioptas *et al.*, 2009] Dimitris Achlioptas, Aaron Clauset, David Kempe, and Cristopher Moore. On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. *Journal of the ACM (JACM)*, 56(4):21, 2009.

[Adamic and Adar, 2003] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[Barabási and Albert, 1999] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[Barabási *et al.*, 2002] Albert-Laszlo Barabási, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and Its Applications*, 311(3-4):590–614, 2002.

[Bunge and Fitzpatrick, 1993] John Bunge and Michael Fitzpatrick. Estimating the number of species: a review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.

[Chen *et al.*, 2016] Lin Chen, Forrest W Crawford, and Amin Karbasi. Seeing the unseen network: Inferring hidden social ties from respondent-driven sampling. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[Dasgupta *et al.*, 2014] Anirban Dasgupta, Ravi Kumar, and Tamas Sarlos. On estimating the average degree. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 795–806. ACM, 2014.

[Eden *et al.*, 2018] Talya Eden, Shweta Jain, Ali Pinar, Dana Ron, and C Seshadhri. Provable and practical approximations for the degree distribution using sublinear graph samples. In *Proceedings of the 2018 World Wide Web Conference*, pages 449–458. International World Wide Web Conferences Steering Committee, 2018.

[Frank, 1980] Ove Frank. Estimation of the number of vertices of different degrees in a graph. *Journal of Statistical Planning and Inference*, 4(1):45–50, 1980.

[Ganguly and Kolaczyk, 2017] Apratim Ganguly and Eric D Kolaczyk. Estimation of vertex degrees in a sampled network. In *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pages 967–974. IEEE, 2017.

[Gao *et al.*, 2016] Jianxi Gao, Baruch Barzel, and Albert-László Barabási. Universal resilience patterns in complex networks. *Nature*, 530(7590):307, 2016.

[Goldreich and Ron, 2008] Oded Goldreich and Dana Ron. Approximating average parameters of graphs. *Random Structures & Algorithms*, 32(4):473–493, 2008.

[Hanley and McNeil, 1983] James A Hanley and Barbara J McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843, 1983.

[Hu and Lau, 2013] Pili Hu and Wing Cheong Lau. A survey and taxonomy of graph sampling. *arXiv preprint arXiv:1308.5865*, 2013.

[Jeong *et al.*, 2001] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41, 2001.

[Jiang *et al.*, 2020] Chunheng Jiang, Jianxi Gao, and Malik Magdon-Ismail. True nonlinear dynamics from incomplete networks. *arXiv preprint arXiv:2001.06722*, 2020.

[Kashtan *et al.*, 2004] Nadav Kashtan, Shalev Itzkovitz, Ron Milo, and Uri Alon. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. *Bioinformatics*, 20(11):1746–1758, 2004.

[Katzir *et al.*, 2011] Liran Katzir, Edo Liberty, and Oren Somekh. Estimating sizes of social networks via biased sampling. In *Proceedings of the 20th International Conference on World Wide Web*, pages 597–606. ACM, 2011.

[Klusowski and Wu, 2018] Jason M. Klusowski and Yihong Wu. Counting motifs with graph sampling. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1966–2011. PMLR, 06–09 Jul 2018.

[Kurant *et al.*, 2012] Maciej Kurant, Carter T Butts, and Athina Markopoulou. Graph size estimation. *arXiv preprint arXiv:1210.0460*, 2012.

[Ribeiro and Towsley, 2012] Bruno Ribeiro and Don Towsley. On the estimation accuracy of degree distributions from graph sampling. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 5240–5247. IEEE, 2012.

[Stumpf and Wiuf, 2005] Michael PH Stumpf and Carsten Wiuf. Sampling properties of random graphs: the degree distribution. *Physical Review E*, 72(3):036118, 2005.

[Stumpf *et al.*, 2008] Michael PH Stumpf, Thomas Thorne, Eric de Silva, Ronald Stewart, Hyeong Jun An, Michael Lappe, and Carsten Wiuf. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964, 2008.

[Szilagyi *et al.*, 2005] Andras Szilagyi, Vera Grimm, Adrián K Arakaki, and Jeffrey Skolnick. Prediction of physical protein–protein interactions. *Physical Biology*, 2(2):S1, 2005.

[Woodward *et al.*, 2010] Guy Woodward, Jonathan P Benstead, Oliver S Beveridge, Julia Blanchard, Thomas Brey, Lee E Brown, Wyatt F Cross, Nikolai Friberg, Thomas C Ings, Ute Jacob, et al. Ecological networks in a changing climate. In *Advances in ecological research*, volume 42, pages 71–138. Elsevier, 2010.

[Zhang *et al.*, 2015] Yaonan Zhang, Eric D Kolaczyk, Bruce D Spencer, et al. Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks. *The Annals of Applied Statistics*, 9(1):166–199, 2015.