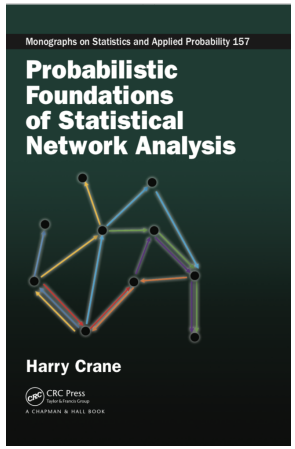
Probabilistic Foundations of Statistical Network Analysis
Chapter 1: Orientation

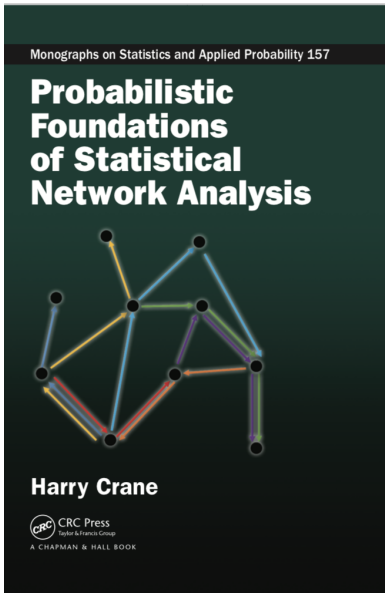
Harry Crane

Based on Chapter 1 of *Probabilistic Foundations of Statistical Network Analysis*



Book website: `http://www.harrycrane.com/networks.html`

Table of Contents



“Networks are everywhere”

In recent years there has been an explosion of *network data* — that is, measurements that are either of or from a system conceptualized as a network — from seemingly all corners of science. *(Kolaczyk, 2009)*

Empirical studies and theoretical modeling of networks have been the subject of a large body of recent research in statistical physics and applied mathematics. *(Newman and Girvan, 2002)*

Networks have in recent years emerged as an invaluable tool for describing and quantifying complex systems in many branches of science. *(Clauset, Moore and Newman, 2008)*

Prompted by the increasing interest in networks in many fields [...]. *(Bickel and Chen, 2009)*

Networks are fast becoming part of the modern statistical landscape. *(Wolfe and Olhede, 2014)*

The rapid increase in the availability and importance of network data [...]. *(Caron and Fox, 2017)*

Network analysis is becoming one of the most active research areas in statistics. *(Gao, Lu and Zhou, 2015)*

Networks are ubiquitous in science. *(Fienberg, 2012)*

“Networks are everywhere”, but ...

- ‘Networks’ arise in a number of ways in a variety of applications. Not a monolithic concept.
- But the prevailing statistical theory of network analysis is mostly based on
  - networks represented as graphs (the ‘networks-as-graphs’ perspective),
  - limited availability of random graph models (stochastic blockmodel, exponential random graph model, graphon models), and
  - models that either ignore sampling outright or assume artificial sampling scheme (e.g., selection sampling or independent vertex sampling).
- Goal of *Probabilistic Foundations*: clarify limitations of current approaches and discuss possible paths forward.

Network analysis as base case of 'complex data science'

- Classical statistics is not about studying sequence of coin tosses, but the theory of coin tossing (i.i.d. Bernoulli trials) lays groundwork for a lot of important ideas in classical statistics (law of large numbers, central limit theorem, etc.).
- Similarly, network analysis is not about studying Facebook, Twitter, or karate club allegiances. Those are merely *applications* of network analysis.
- Network analysis is the 'base case' of modern complex data analysis.

	Classical statistics	**Complex data analysis**
Canonical setting	Coin tossing/Bernoulli trials	Networks/Complex systems
Data structure	'flat' (sequences/sets)	'deep', structured
Data behavior	regular (i.i.d., stationary)	irregular (hetergeneous, fat tails)
Major insights	SLLN, CLT, etc.	'Scale-free', little theory
Modern challenge	**Computational challenges** 'Big data' (flat, but bigger)	**Conceptual challenges** structure *as* data

$$\underbrace{\text{coin tossing : unstructured data}}_{\textit{classical statistics}} :: \underbrace{\text{network analysis : complex, structured data}}_{\textit{modern/future 'data science'}}$$

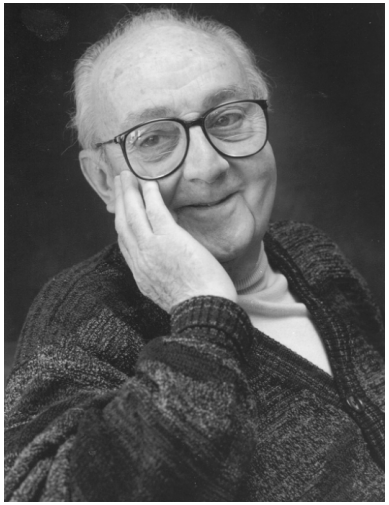
Modeling as foundation of Network Analysis

1. Modeling is the act of imposing structure on the data (and thus on the world).
- In classical statistics the data is assumed to have little structure (sets, tables, arrays).
- In modern applications, the data has complex, heterogeneous structure (dependence, interactions, relations).
- In network analysis, the structure *is* the data.
- Future of data science: framework for analyzing data with built-in, complex structure.

2. For most interesting applications, existing network models (SBM, ERGM, graphon) do not live up to their names as 'models'. (Chapters 2, 6–8)
3. Despite limitations to these models, their appearance in the theoretical and applied statistics literature is pronounced. No desire to continue this trend here.
4. De-emphasis of modeling in classical statistics courses.
- Boxian trope, "All models are wrong, but some are useful."

All models are wrong ...

All models are wrong, but some are useful.



George Box (1919–2013)

What does it mean for a model to be 'wrong'? What makes a model 'useful'?

- Network analysis: Answer not straightforward $\rightarrow$ will explain why most current network models are not useful in most situations.
- Chapter 5: modeling paradigm (Crane–Dempsey) proposed to make future theoretical research in network analysis more useful.

‘Networks-as-graphs’ perspective

Networks are not graphs.

- Graph: mathematical object consisting of vertices V and edges $E \subseteq V \times V$. (More general forms with multiple edges, weighted edges, multiple layers, etc.)
- Network: abstract concept referring to system of interrelated entities. Not well-defined concept. Is vague/amorphous.

- Sometimes reasonable to represent networks mathematically as graphs, but not always.
- Progress in network science: disentangle notions of *graph* (mathematical) and *network* (conceptual).
- Need other ways to represent networks. Still limited options. Crane–Dempsey edge exchangeable models (Chapters 9–10) are one recent proposal.

Representing and labeling networks

- Data analysis is like trying to discern nature of a large, complex object in a dark room using only a small flashlight.
- The ‘angle’ from which the light shines on the object determines which aspects of the object are visible.
- In data analysis, this ‘angle’ refers to data representation and the way the data is obtained (sampled, generated).
- Figure below shows 3 ways of representing the ‘same’ network. Are these three representations the same?

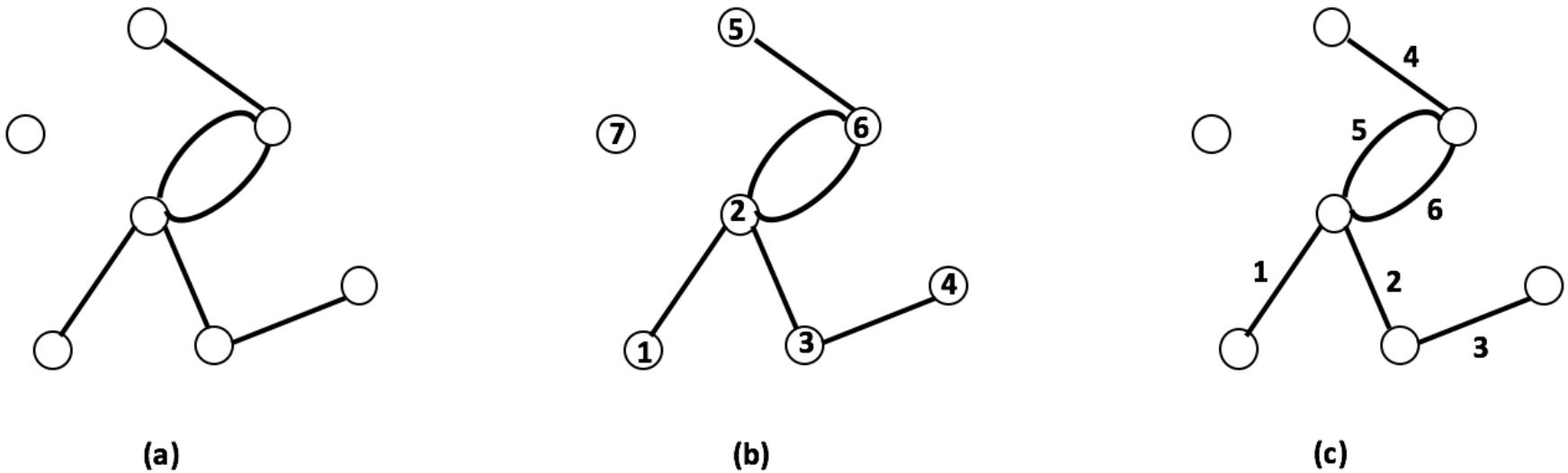


Figure: Figure 1.1 of *Probabilistic Foundations of Statistical Network Analysis*.

The role of context in network analysis

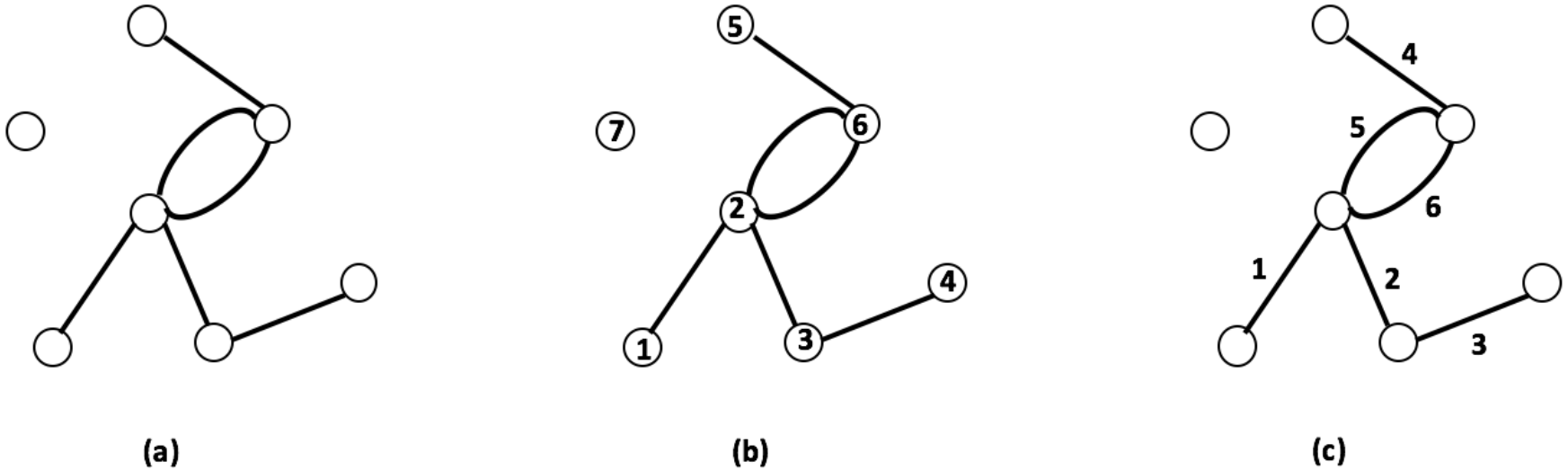


Figure: Figure 1.1 of *Probabilistic Foundations of Statistical Network Analysis*.

- Do the three networks above represent the same network? Perhaps.
- Assuming they do, are they the same *representation* of that network? Of course not.
- Part (a): 'shape' of the network (i.e., 'unlabeled' network).
- Part (b): vertex-labeled graph representation.
- Part (c): edge-labeled graph representation.
- What is the difference and why choose one over the other? Depends on context of the application.

‘Being useful’ = ‘Making sense’ $\Rightarrow$ Coherence

A statistical model has two primary components:

- *Descriptive component*: consists of family of candidate probability distributions for describing variability in the observed data.
- *Inferential component*: explains how the observed data fits into a larger context.

Example: Let $X_1, \ldots, X_n$ be i.i.d. random integers and let Y be obtained by size-biased sampling from $X_1, \ldots, X_n$. Then the ‘size-biased sampling’ puts Y in the context of $X_1, \ldots, X_n$.

- Induces distribution of Y:

$$\Pr(Y = k) \propto k \Pr(X_i = k), \quad k \geq 1.$$

Boxian proverb: “All models are wrong, but some are useful.”

- Call a model ‘coherent’ if its inferences can be interpreted in a single (coherent) context.
- Then a model is ‘useful’ only if it is coherent.
- More details in Chapter 5.

Examples of network data

Whenever possible, I motivate different models by a simple example.

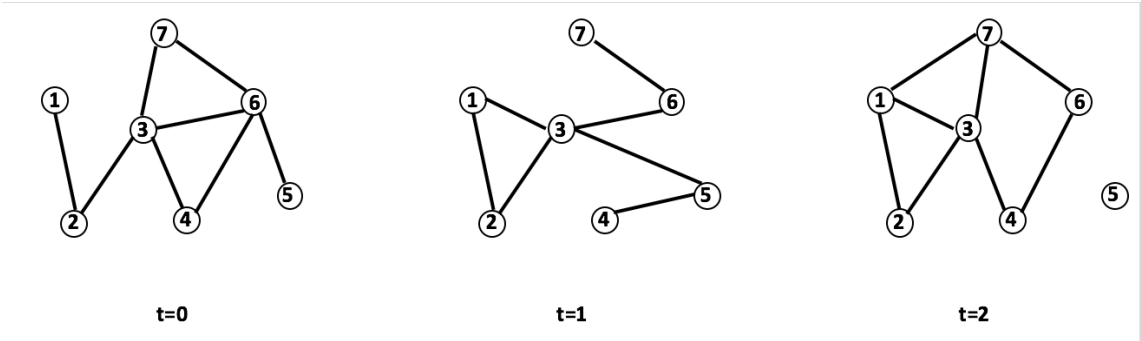
- Caveat: For the most part, these examples are not interesting on their own, and are offered here only to illustrate how basic principles of network analysis arise in practice.

Common examples:

- Internet network
- Social networks (Facebook, Twitter)
- Karate club network
- Enron email corpus
- Collaboration networks
- Blockchain and cryptocurrency networks
- Brain networks
- Gene regulatory networks
- Telecommunications networks
- etc.

Major Open Questions (Section 1.7)

1. Modeling and analyzing ‘complex’ networks (Chapters 4, 6–10)
   - Many real-world networks are ‘sparse’ (i.e., few edges relative to vertices) and ‘scale-free’ (i.e., power law degree distribution).
   - In one sense, poorly connected (because sparse), but in another sense well-connected (because of heterogeneous, ‘complex’ edge patterns due to scale-freeness).
   - Prevailing approaches to statistical modeling (SBM, graphon, ERGM) are models for homogeneous (dense, light-tailed) networks. Do not account for complexity.
2. Network sampling (Chapters 2, 3, 5–10)
   - Many networks are sampled, but current statistical network theory doesn’t account for network sampling.
3. Network dynamics (Chapter 11)
   - How to analyze networks that vary with time?



Tackle some of these questions throughout the book. Many open challenges remain.

Toward a Probabilistic Foundation for SNA

- Chapter 1: Vision for network analysis as framework for ‘complex data science’.
- So far: Vision has not been fulfilled.
- Current focus of the field on random graphs, community detection, graphon estimation, minimax optimal rates, power law properties, etc. miss the point.
- Goal of this book:
  1. Emphasize the need for a new perspective on network analysis.
  2. Enlist readers in fulfilling the vision of network analysis as framework for complex data science.

Based on Chapter 1 of Probabilistic Foundations of Statistical Network Analysis