

Complex-Network Modelling and Inference

Lecture 22: Network Topology Measurement

Matthew Roughan

`<matthew.roughan@adelaide.edu.au>`

https://roughan.info/notes/Network_Modelling/

School of Mathematical Sciences,
University of Adelaide

January 14, 2025

Section 1

Network Topology Measurement

Motivation

- Lots of places where we need to measure the network
- Some are surprising
 - ▶ surely network managers know what their network looks like?

How Motivations Affect Techniques

- Scientific – networks are of great scientific interest
- Adversarial – competitors and customers want information that they might not be naturally privy to
- Managerial – the network should be the database of record

Scientists and adversaries are often in essentially the same position from the point of view of techniques available to them.

What Network

Even for something as well defined as the Internet this needs to be considered carefully.

- layer 1 - physical level
- layer 2 - link/switch level
- layer 3
 - ▶ IP router level
 - ▶ PoP level
 - ▶ AS-level
- Application level (web, P2P, Social nets)

Physical vs virtual matters

What data?

Might not just want connectivity

- link capacities
- link length (or even its entire physical path)
- node location
- node type
 - ▶ brand, version, ...

Routing

- We'd like to learn what paths are actually being used
- We might even like to learn how they were determined
 - ▶ e.g., if shortest paths, what are the link weights?

Generic Approaches

- node data:
 - ▶ each node “tells” you its links
- edge data:
 - ▶ each edge “tells” you the nodes it connects
- path data:
 - ▶ measure a set of paths
 - ▶ usually ordered, but can be unordered (co-occurrences)
- inverse problems:
 - ▶ we learn some data from which we infer edges/nodes
 - ▶ more on this later

Node data

Examples:

- Surveys (e.g., you fill out a form, and list your friends)
- Fetch facebook or WWW pages, and parse them for connections
- Read scientific paper
 - ▶ get co-authors
 - ▶ get citations
- ???

Node data

- Pros:
 - ▶ direct
 - ▶ simple
 - ▶ often informative about node properties
- Cons:
 - ▶ relies on co-operation of nodes
 - ★ sometimes a node is “dumb”, e.g., a vole
 - ★ sometimes a node won't co-operate with outsiders
 - ▶ sometimes we don't start with a list of nodes, or any way to find them

Edge data

Examples:

- Twitter tweets give you an implicit edge
 - ▶ more generally, we can sometimes observe “traffic”
- Observe contacts, e.g., vole study
- Biochemistry: we perform chemical experiments
- Parse movie script
 - ▶ edges created by shared scenes
 - ▶ OK, maybe this is a little bit of a stretch

Edge data

- Pros:
 - ▶ gives you the nodes (except for degree zero nodes), which is useful when these are hard to access
- Cons:
 - ▶ not an easy approach in many cases as edges are even more likely to be “dumb”
 - ▶ many more edges than nodes (in most cases)
 - ▶ disconnected nodes are invisible

Path data

Examples

- co-occurrence data
 - ▶ co-excitation of biochemicals in a cell corresponding to a signalling pathway
 - ▶ activated genes
 - ▶ switches activated by a telephone call
 - ▶ co-cited academic papers
 - ▶ fMRI images of brain
- path data (ordered co-occurrences)
 - ▶ traceroute in the Internet
 - ▶ GPS tracking of taxis
 - ▶ Milgram experiment
- tree data
 - ▶ multicast tree

Path data

- Pros:
 - ▶ similar advantages to edge measurements
 - ▶ get to see routes as well as topology
- Cons:
 - ▶ errors and missing data are common
 - ▶ ordering may be hard to ascertain, and without ordering we have to do some sort of inference
 - ▶ unique labelling may not work
 - ▶ can't see zero-degree nodes
 - ▶ can't see "latent" paths
 - ★ backup pathways in the Internet
 - ★ routes that taxis don't use

Node Data Example: Web Crawling

- Challenge is the size and dynamic nature of problem
 - ▶ takes some time to crawl the whole thing
 - ▶ its changing as you crawl it
 - ▶ how do you make sure you have seen it all
 - ★ what about disconnected components
- Social network crawling has similar problems

Node Data Example: Surveys

- Traditional method for collecting data on social networks
- Problems:
 - ▶ Limited number of responses:
 - ★ issues for samples of a graph
 - ★ biases in sample?
 - ▶ Responses aren't always accurate
 - ★ people lie

Path Data Example: Traceroute

- developed by Van Jacobsen around 1988 [Smi, Jac04]
- Time-To-Live (TTL) [Ste94, rfc81, Bak93] field of an IP packet
 - ▶ decremented at each hop
 - ▶ when gets to zero, router at which this happens stops forwarding the packet (hence terminating any loops) and generates an ICMP “Time Exceeded” message that is returned to the source.
- Send out a series of packets
 - ▶ (1) TTL=1
 - ▶ (2) TTL=2
 - ▶ ⋮

Example

traceroute slashdot.org is given below:

traceroute to slashdot.org (66.35.250.150), 30 hops max, 38 byte packets

```
 1  129.127.5.254  0.188ms  0.393ms  0.296ms
 2  129.127.254.17  0.236ms  0.352ms  0.411ms
 3  129.127.254.19  0.499ms  0.543ms  0.512ms
 4  129.127.254.190 1.766ms  0.548ms  *
 5  192.43.227.19  1.912ms  0.788ms  0.719ms
 6  138.44.192.17  2.077ms  0.777ms  0.740ms
 7  113.197.15.28  11.812ms 9.728ms  9.727ms
 8  113.197.15.8   22.416ms 21.224ms 21.196ms
 9  113.197.15.2   23.588ms 22.296ms 22.326ms
10  113.197.15.143 24.034ms 22.673ms 22.717ms
11  202.158.194.121 166.149ms 164.958ms 165.069ms
12  64.125.193.129 166.253ms 164.991ms 164.920ms
13  63.146.26.81   177.170ms 175.942ms 175.858ms
14  63.235.40.210  177.114ms 175.947ms 175.900ms
15  206.28.98.174  218.970ms 217.644ms 217.006ms
16  206.28.96.185  217.750ms 215.752ms 216.883ms
17  204.70.196.230 218.167ms 220.271ms 219.437ms
    ⋮
```

Problem 1: non-atomicity

- Traceroute is made up of a series of measurements over time
- If routing changes during measurement, its results are meaningless
- changes due to
 - ▶ network change or failure
 - ▶ load balancing

Problem 2: aliasing

- Responding router uses an IP address
 - ▶ routers have IP address for each interface, plus loopback
 - ▶ response IP can vary
- How do you put a series of paths together into a topology when nodes don't have unique labels?

Problem 3: missing data

- Traceroute requires control over a source
 - ▶ limited number of public traceroute servers
 - ▶ limited number of viewpoints
 - ▶ only see forward path
- Some places deliberately prevent it from working
 - ▶ routers don't respond
 - ▶ responses are filtered
 - ▶ incoming ICMP or UDP are filtered

Missing data is not “missing at random”

Traceroute summary

- Traceroute packets may be blocked or dropped, leaving some areas of the network blank
- Traceroute provides paths without unique node identifiers
- Traceroute can only see utilised links
- Traceroute is non-atomic – it is made up of a series of measurements over some time interval. If the network changes during this interval, the results are garbage.
- Traceroute can only see a forward path from a measurement station. There are a limited number of these, and so it is often impossible to scan a complete network.
- Traceroute can only see IP hops
- Traceroutes cannot see, and are easily fooled by link-layer technology, such as MPLS (Multi-Protocol Label Switching).

Path Data Example: Social Networks

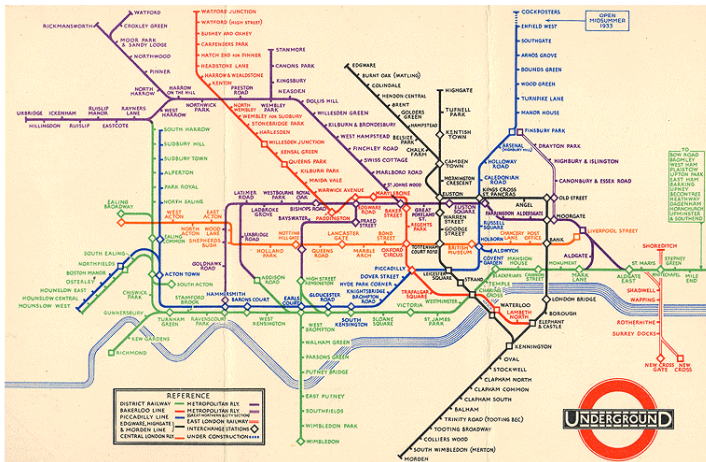
- e.g. Stanley Milgram [**Mil67**] experiment
- See path from end-to-end
- Highly sampled part of much larger network

UPDATES!!

- “An Experimental Study of Search in Global Social Networks,”
Dodds, Muhamad and Watts, Science, Vol. 301, 2003.
 - ▶ Average over *completed* chains 4.05
 - ▶ But biased because shorter chains more likely to be completed
 - ▶ Reconstructed median = 7
- **How small is the world, really?**, Watts, 2016
 - ▶ Argues that hard to make it any smaller

Path Data Example: London Underground

1933



Beck's famous London "tube" map

<http://www.bbc.co.uk/news/uk-england-london-20943525>

Path Data Example: Wikispeedia

<http://snap.stanford.edu/data/wikispeedia.html>

Human navigation paths on Wikipedia: Wikispeedia, users are asked to navigate from a given source to a given target article, by only clicking (condensed version of) Wikipedia links.

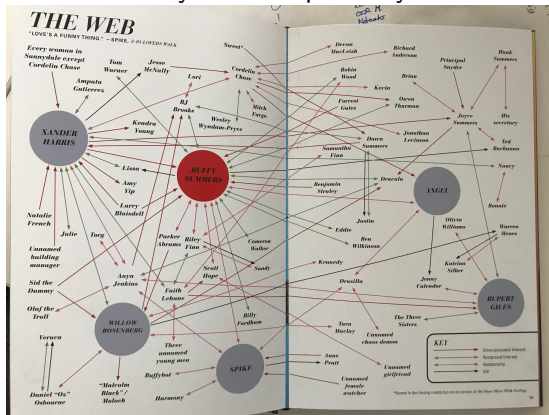
Unordered Path data

- Network Inference from CoOccurrences (NICO) [BNR11, RFN08]
- Assume we can measure the set of simultaneously active nodes or edges, but not the ordering
 - ▶ how can we get path/graph data
 - ▶ turns out to be a nice inference problem, using the EM algorithm
- Used, for instance,
 - ▶ in telephone network where timing data isn't accurate enough to determine order
 - ▶ in determining metabolic networks: we can see which proteins are active, but not which affects which

Co-occurrence more generally

- Networks in narratives: literature, TV, movies
 - ▶ e.g., character networks

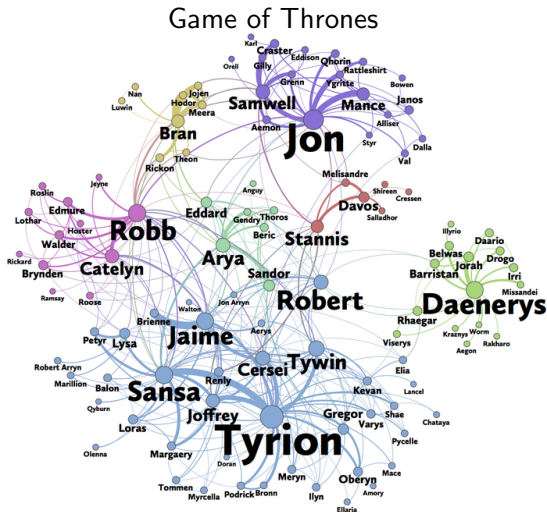
Buffy the Vampire Slayer



From "Buffy the Vampire Slayer: Slayer Stats", Guerrier and O'Brien

Co-occurrence more generally

- Networks in narratives: literature, TV, movies
 - ▶ e.g., character networks



From "Network of Thrones", <https://networkofthrones.wordpress.com/>

Co-occurrence more generally

- Networks in narratives: literature, TV, movies
 - ▶ e.g., character networks
- Would like to model characters that interact, but that is hard
 - ▶ instead look at characters that are in the same “chunk”
 - ★ are in the same page/section/chapter
 - ★ are in the same scene/episode

Edges connect characters that appear close together in the narrative

- Problems:
 - ▶ We are measuring a *proxy* of the underlying network
 - ★ false positive links: essentially a “scene” becomes a clique, but not everyone in a scene interacts
 - ★ false negatives: too short a chunk, and you miss some real interactions







Q: how long a chunk should be used to construct a clique?

- Proxy network measurements are COMMON, but often pitched as “THE” network

Lessons

- You **HAVE** to think about how you will measure a network
- Not all methods are created equal
 - ▶ some are more work, but will give better results
 - ▶ idea methods are sometimes impractical
- Measurements have artifacts, errors, missing data, ...
 - ▶ understand them!
- Always define **EXACTLY** what your network is
 - ▶ what are the nodes
 - ▶ what are the edges (how were they constructed and what do they mean?)

Further reading I

-  F. Baker, *Requirements for IP version 4 routers*, IETF, Network Working Group, Request for Comments: 1812, July 1993.
-  Laura Balzano, Robert Nowak, and Matthew Roughan, *On the success of network inference using a markov routing model*, International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 2011.
-  Van Jacobson, *Traceroute*, <ftp://ftp.ee.lbl.gov/traceroute.tar.gz>, 1989-04.
-  Stanley Milgram, *The small world problem*, *Psychology Today* **1** (1967), no. 1, 60–67.
-  *Internet Protocol*, IETF RFC 791, September 1981.
-  M. G. Rabbat, M. A. T. Figueiredo, and R. D. Nowak, *Network inference from co-occurrences*, *IEEE Transactions on Information Theory* **54** (2008), no. 9, 4053–4068.

Further reading II



Craig Smith, *Traceroute - whitepaper*,

<http://www.informatik.uni-trier.de/~smith/networks/tspec.html>.



W. Richard Stevens, *TCP/IP illustrated, volume 1*, Addison-Wesley Publishing Company, 1994.