Communities, hierarchies, and cores: finding large-scale structure in complex networks

Dan Larremore
BioFrontiers Institute & Dept. of Computer Science

Nov 10, 2017
Computational Social Science Initiative Methods Workshop

University of Colorado **Boulder**

daniel.larremore@colorado.edu
@danlarremore

Of course: PDF of slides available after I fix any typos tonight.

The Plan

	Goals	Datasets
9:00 - 12:00 Workshop	**Community Structure**	**Faculty Hiring**
12:00 - 1:30 Lunch	**Linear rankings & hierarchies**	**Malaria Parasites**
1:30 - 4:30 Workshop	**Practice**	**[Your Fav. Dataset Here]**

All times listed ± n minutes. n is some positive integer.
The point is to learn, not to make it through all the slides.

Simplicity is a great virtue but it requires hard work to achieve it and education to appreciate it. And to make matters worse: complexity sells better.

E. W. Dijkstra

Take this two ways:

The Cynic: Pictures of networks can be *really cool* but our goal is to do good science, not make pretty pictures.

The Scientist: The most beautiful science is when we *correctly* simplify [the simple parts of] a complex system.

Networks vocabulary whirlwind

Network. Graph. Multigraph. Loopy multigraph.

Edge. Link. Tie. Node. Vertex. Ego. Alter.

Directed vs Undirected.

Unweighted vs Weighted.

Adjacency Matrix. Degree.

Bipartite. Projection.

Metadata.

Partition. Embedding.

What other words have you heard?

What do we mean by “large-scale structure” ?

Structure is what makes data different from noise.
It’s what makes a network different from a random graph.

Networks can be too large and complex to be adequately summarized by a few scalars (node or link count, average degree). And yet, they are too large and complex to be taken in all at once.

Understanding what the network *means* requires that we highlight key structures.

Searching for large-scale structures in a network reflects a belief that in all the complexity there are patterns that make the network *less* complicated.

We define these structures—models, really—to compress complex networks.

Understanding, not a list of parts and dimensions

Finding large-scale structures is the same as anything else:

a simplified model of something very complicated.

We want to know how it all fits together.

Many uses for the same techniques. cf regression

Treat the network like a system:

Extrapolation. Make predictions for as-yet unseen nodes (in “space” or time).
Interpolation. Identify missing links.
Generalization. Nodes of this type are like others of the same type.

Treat the network like an artifact:

Mechanisms. How did this network arise? What rules governed its assembly?
Explanations. Coarse-graining or compression.

Treat the network like a means to an end; an intermediate data structure:

Useful division. Need groups so that we can assign treatments in an A/B test.
Simplification. Downstream regression model needs ranks or groups.

Community structure

Homophily & assortative mixing

like links with *like*

Assortativity coefficient *r* measures extent of homophily.

Newman, Phys. Rev. E 67, 026126 (2003).

Homophily & assortative mixing

like links with *like*

Assortativity coefficient *r* measures extent of homophily.

Three types:

- scalar attributes
- vertex degrees
- categorical variables

Newman, Phys. Rev. E 67, 026126 (2003).

Homophily & assortative mixing

like links with *like*

Assortativity coefficient *r* measures extent of homophily.

scalar attributes:

compute a Pearson correlation over edges.

start with the mean of c across ties:

$$\mu = \frac{1}{2m} \sum_i \sum_j A_{ij} c_i = \frac{1}{2m} \sum_i k_i c_i$$

Newman, Phys. Rev. E 67, 026126 (2003).

Homophily & assortative mixing

like links with *like*

Assortativity coefficient *r* measures extent of homophily.

compute covariance:

$$\mathrm{cov}(c_i, c_j) = \frac{\sum_{ij} A_{ij}(c_i - \mu)(c_j - \mu)}{\sum_{ij} A_{ij}}$$

$$= \frac{1}{2m} \sum_{ij} A_{ij} c_i c_j - \mu^2$$

$$= \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) c_i c_j$$

use $\mu = \frac{1}{2m} \sum_i k_i c_i$

Newman, Phys. Rev. E 67, 026126 (2003).

Homophily & assortative mixing

like links with *like*

Assortativity coefficient (scalar).

$$r = \frac{cov(c_i, c_j)}{var(c_i, c_j)}$$

$$r = \frac{\sum_{ij}(A_{ij} - k_i k_j / 2m) c_i c_j}{\sum_{ij} k_i \delta_{ij} - k_i k_j / 2m}$$

"it's that easy!"

$$-1 \leq r \leq 1$$

Newman, Phys. Rev. E 67, 026126 (2003).

Homophily & assortative mixing

like links with *like*

degree: just another scalar.

(very well studied!)

Newman, Phys. Rev. E 67, 026126 (2003).

Homophily & assortative mixing

like links with *like*

categories: not just another scalar.

Imagine that *c* are labels, categories.

let e_{rs} be the fraction of edges between nodes of type r and s .

$$\sum_{rs} e_{rs} = 1 \qquad \sum_{r} e_{rs} = a_r \qquad \sum_{s} e_{rs} = b_s$$

then, $r = \dfrac{\sum_r e_{rr} - \sum_r a_r b_r}{1 - \sum_r a_r b_r}$

consider: if there were *only* edges between nodes of the same type…

Homophily & assortative mixing

like links with *like*

It is often convenient to write this in a different way, and call it *Q.*

Same principle: what fraction of edges fall between nodes of the same community?*

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{ij}$$

*compared to what we'd expect if the network were random!

Aside: the birth of null models & chance sociograms

Who shall survive? Moreno, 1936

Moreno wondered if there were structural explanations for why certain young girls were running away from the school, and thought that sociographic analysis might hold an answer.

chance sociograms

SIAM Review: Configuring random graph models with fixed degree sequences. http://arxiv.org/abs/1608.00607

The Book: http://www.asgpp.org/docs/wss/Book%20VI/index.html

Johan Ugander's Post: https://jugander.wordpress.com/2014/08/07/computational-perspectives-on-large-scale-social-networks-a-brief-history/

Practice makes the master

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{b_i, b_j}$$

$$Q = \sum_r e_{rr} - a_r^2$$

labeling 1	red	blue
red	3/7	1/14
blue	1/14	3/7

$Q_1 = 5/14 = 0.357$

labeling 2	red	blue
red	4/7	2/14
blue	2/14	1/7

$Q = 6/49 = 0.122$

Aaron Clauset: http://tuvalu.santafe.edu/~aaronc/courses/5352/fall2013/csci5352_2013_L5.pdf

Modularity

Modularity is easily *the* most popular method for community detection. But why?

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{b_i, b_j}$$

Why is this more powerful than simply a measure of correlation over node labels?

Community structure in social and biological networks
M **Girvan**, MEJ **Newman** - Proceedings of the national …, 2002 - National Acad Sciences
Abstract A number of recent studies have focused on the statistical properties of networked systems such as social networks and the Worldwide Web. Researchers have concentrated particularly on a few properties that seem to be common to many networks: the small-world
☆ ❝❝ Cited by 10319 😱 ted articles All 54 versions

Finding and evaluating community structure in networks
MEJ **Newman**, M **Girvan** - Physical review E, 2004 - APS
Abstract We propose and study a set of algorithms for discovering community structure in networks—natural divisions of network nodes into densely connected subgroups. Our algorithms all share two definitive features: first, they involve iterative removal of edges from
☆ ❝❝ Cited by 9247 😱 ted articles All 44 versions

Girvan & Newman, 2002. Community structure in social and biological networks. PNAS 99, 2002.

Key: let's reverse our thinking of what Q does

Don't use Q to compute correlation of some given labels.

Instead, fiddle with the labels and see how you can maximize Q!

Now, we have a CS problem: how do
you search the space of partitions?

(This space is really big!)

How would *you* do it?

People like modularity. Why?

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{b_i, b_j}$$

- Intuitive
- Works for weighted and unweighted networks.
- Corresponds to our social network ideas of what (cohesive) communities are.
  - Automatically choose *k,* the number of groups.
  - Rapid approximate solutions.
  - Follows the usual trajectory: idea, demonstration, optimization.
  - Fun customizations:
    - Resolution parameter to “zoom in” and “zoom out.”
    - Find the clusters. Then cluster the clusters. Then cluster those clusters…
    - Directed. Bipartite.

$$Q = \frac{1}{m} \sum_{ij} \left(A_{ij} - \frac{k_i^{\text{out}} k_j^{\text{in}}}{m} \right) \delta_{b_i, b_j} \qquad Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \gamma \frac{k_i k_j}{2m} \right) \delta_{b_i, b_j}$$

Why aren’t we done here?

Physicists like to minimize things because rocks fall.
— Cris Moore

We can always maximize Q to find a partition, but is it meaningful?

Random networks

Turns out, you can find high-modularity partitions in random networks.

Structure is what makes data different from noise.
It's what makes a network different from a random graph.

We prefer that our methods fail gracefully, and tell us when they fail. (cf R^2)
Alternative perspective: maybe you *want* to find clusters in randomness?

Guimera, Sales-Pardo, Amaral. "Modularity from fluctuations in random graphs and complex networks." Physical Review E 70.2 (2004): 025101.

A basket of degenerates

Lots of different but nearly-as-good partitions.
The optimization landscape is *degenerate*.

Unintuitive behavior: find the communities in a chain of cliques and you get pairs of cliques…not the cliques themselves!

Good, de Montjoye, Clauset. Performance of modularity maximization in practical contexts. Phys. Rev. E 81, 046106 (2010).

Q is restricted to *assortative* community structure

The zoo of possible structures is diverse and interesting!
Build intuition: what do these networks look like?

Assortative

Disassortative

Ordered

Core-periphery

If only there were a more flexible model . . .

Break

Here is my favorite paper of all time:

JMLR: Workshop and Conference Proceedings 27:65–79, 2012 Workshop on Unsupervised and Transfer Learning

Clustering: Science or Art?

Ulrike von Luxburg ULRIKE.LUXBURG@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems, Tübingen, Germany

Robert C. Williamson BOB.WILLIAMSON@ANU.EDU.AU
Australian National University and NICTA, Canberra ACT 0200, Australia

Isabelle Guyon ISABELLE@CLOPINET.COM
ClopiNet, 955 Creston Road, Berkeley, CA 94708, USA

http://proceedings.mlr.press/v27/luxburg12a/luxburg12a.pdf

Task 1: github

git **clone**
git **pull**
git **status**
git **add**
git **commit -m “stuff”**
git **push**

Task: clone my data repo from a 2013 PLoS Comp Bio paper

Beyond assortativity: block models

What do these have in common?

Assortative

Disassortative

Ordered

Core-periphery

Beyond assortativity: block models

What do these have in common?

Assortative

Disassortative

Ordered

Core-periphery

Nodes are in groups with other nodes that *connect to other groups in similar ways.*

Key idea: all nodes in a group are *stochastically equivalent*.

Generative model approach

Generate the structure you wish to infer.

We like generative models because they open the door to inference:

$$\text{GM + parameters} \xrightarrow{\text{stochastic draw}} \text{Data}$$

$$\text{GM + parameters} \xleftarrow{\text{inference}} \text{Data}$$

In other words: let's write down a recipe for generating block structure.

The stochastic block model

GM + parameters $\xrightarrow{\text{stochastic draw}}$ Data

Assign each node to one of *B* blocks.

b_i

Let the probability that two nodes connect depend *only* on their blocks:

$$Pr(A_{ij}|b_i, b_j) = \omega_{b_i,b_j}$$

Then we can choose the matrix ω to have whatever structure we want!

Assortative

Disassortative

Ordered

Core-periphery

SBM inference

GM + parameters $\xleftarrow{\text{inference}}$ Data

no more math on slides 😭

(boringness prevention intervention)

Karrer, Newman. Stochastic blockmodels and community structure in networks. Phys. Rev. E 83, 016107 (2011).

If you're tuning in at home…

The problem with parameterized models…

You have to choose their parameters!

How shall we choose *B*, the number of blocks?

Hint: can't just use the likelihood:

Why? If we place each node in its own community, we can get Likelihood=1.

[Actually, this wouldn't model the data at all: it would *memorize* it.]

We need a way to penalize the complexity of the model. Any ideas?

Task 2: Game Interlude

Suppose that you actually like Twitter.

You have 140 characters.

Explain your current grant to me, in as much detail as you can manage. Feel free to use your best space-saving and character-frugality approaches.

Email them daniel.larremore@colorado.edu

https://wordcounter.net/

Description length & Occam's razor

The description length of a message is:
bits required to send the compressed message + # bits in encoding scheme.

Occam's razor: among all possible explanation for a phenomenon, choose the simplest one.

The stochastic block model also has a description length:

$$\boxed{\Sigma} = \boxed{\mathcal{S}} + \boxed{\mathcal{L}}$$

description length = entropy of data, given the model (fit SBM) + entropy of model

Consider the original problem:
what happens to this equation when I increase the number of blocks B?

Peixoto. Entropy of stochastic blockmodel ensembles. Phys. Rev. E 85, 056122 (2012).
Peixoto. Parsimonious Module Inference in Large Networks. Phys. Rev. Lett. 110, 148701 (2013).

MDL criterion suggests an algorithm:

Fit the SBM with 1 block and record the DL.
Fit the SBM with 2 blocks and record the DL.
…
when the DL starts to increase, go back a step and stop.*

Bonus: what happens if I trick you and give you a *random network*?

MDL approach will tell you: this is *better modeled by* a random network.

*Actually, use something clever, like Golden Ratio / Fibonacci search
Press et al. Numerical Recipes: The Art of Scientific Computing, (Cambridge University Press, Cambridge, England, 2007), 3rd ed.

So how does the search part work?

Markov-chain Monte Carlo:

Wander from partition to partition by proposing to take a node from one group and move it to a new group.

If it's a good move, do it.

If it's a bad move, maybe do it—depends on how bad it is.

See documentation for graph-tool. https://graph-tool.skewed.de

Does it work?

Adamic & Glance mapped the link structure of political blogs in 2004.

Karrer & Newman used this network as a testbed for community detection using the SBM.

What does this say about the process that may be generating (or pruning?) the links in this network?

Division in political blog network.

Karrer & Newman PRE 2011. https://arxiv.org/abs/1008.3926

Adamic & Glance KDD 2005. https://dl.acm.org/citation.cfm?id=1134277

Does it work?

In bipartite networks, we *know* the major split in the data already.

Methodologically, we found that exploiting this split improved speed and quality of the partitions we found.

Scientifically, this opened new directions to analyze (and understand) evolutionary constraints on malaria parasites.

Genes & substrings,
malaria immune evasion

Code by Tzu-Chi Yen https://github.com/junipertcy/det_k_bisbm.
Figure http://danlarremore.com/webweb/

Larremore, Clauset, Buckee, *PLoS Comp Biol*, 2013.
Larremore, Clauset, Jacobs, *Physical Review E*, 2014.

Does it work?

In bipartite networks, we *know* the major split in the data already.

Methodologically, we found that exploiting this split improved speed and quality of the partitions we found.

Scientifically, this opened new directions to analyze (and understand) evolutionary constraints on malaria parasites.

Genes & substrings,
malaria immune evasion

Larremore, Clauset, Buckee, *PLoS Comp Biol*, 2013.
Larremore, Clauset, Jacobs, *Physical Review E*, 2014.

Does it work?

C. elegans neuronal network.

302 neurons, completely mapped. The neurons do not fire action potentials, and do not express any voltage-gated ion channels.

Note the different layout…

C. elegans 1st *multicellular* genome.1998. http://science.sciencemag.org/content/282/5396/2012.
Bacteriophage 1st genome: https://www.nature.com/articles/265687a0

A good alternative? cross-validation via link prediction

Select *B* by choosing the model that makes the best predictions.

Perform *k*-fold cross validation:

1. Divide the edges of the network into *k* groups, called folds.
2. Hide one of the folds (the “test set”)
3. Fit each SBM to the remaining k-1 folds (the “training set”), varying B.
4. Test the ability of the fitted models to predict the hidden test data.
5. Switch which fold is “test” and which are “training” and repeat.

Choose the B with the highest performance on link prediction over all k folds.

Note: this is exactly how we write [math] exams, minus the “repeat”.

Degree Correction

It turns out that I hid something from you before about the stochastic block model.

Rather than generate block-wise random matrices, we want to be able to do so but *also* choose the degrees of the nodes.

Generative model perspective: if we can't generate communities/networks with broad degree distributions, we can't find those communities in real data.

What can be done? **Correct for (or *control* for) the degrees of the nodes.**

Degree correction

no more math on slides 😭

(boringness prevention intervention)

Karrer, Newman. Stochastic blockmodels and community structure in networks. Phys. Rev. E 83, 016107 (2011).

If you're tuning in at home…

SBM vs “DC” SBM

(a) Without degree-correction

(b) With degree-correction

The degree-corrected stochastic block model (right) finds communities which correspond to known metadata.

Karrer, Newman. Stochastic blockmodels and community structure in networks. Phys. Rev. E 83, 016107 (2011).
Adamic, Glance. The political blogosphere and the 2004 US election: divided they blog. 36–43 (2005).

Task 3: graph-tool

If you don’t have python and graph-tool running,
not to worry! Get a server login from me.

Server note:
https://54.244.78.119

* disappears after today
* set up by CU BioFrontiers IT wizards
* running on AWS so we get the shady certificate message.
    * -> proceed anyway

Advanced topic 1: hierarchical communities

Don't minimize the description length of the data and the model…

Model the model as well. Why?

If we compress the model, we can afford a bigger model, but a lower overall cost.

Except now, the description length includes two models. Or three? Or?

Peixoto. Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. Phys. Rev. X 4, 011047 (2014).

Advanced topic 1: hierarchical communities

Peixoto. Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. Phys. Rev. X 4, 011047 (2014).

Advanced topic 2: mixed-membership

Nodes are often pulled between communities. (Or in real social systems, individuals belong to multiple groups.)

“Mixed membership” models allow for that, by assigning *links* to groups, and assigning nodes to groups based on their links.

Ball, B., Karrer, B. & Newman, M. E. J. Efficient and principled method for detecting communities in networks. Phys. Rev. E 84, 036103 (2011).

Advanced Topic 3: multilayer networks

In single-layer networks:
nodes and edges

In multilayer networks:
nodes, edges, and layers

edges: different types of relationships
layers: each layer contains all edges of one type
nodes: same nodes in each layer

Multilayer network: air travel

traditional: booking with airline

disrupted: booking with kayak, expedia, etc

Multilayer network: community structure?

three key approaches:

1. **Non-generative**: modularity maximization; vary inter-layer strength.
   Mucha et al *Science* 2010. http://science.sciencemag.org/content/328/5980/876

2. **Generative**: SBM for each layer, but jointly model layers whenever their structures are sufficiently similar.
   Peixoto, T. P. Phys. Rev. E 92, 042807–15 (2015).

3. **Generative**: SBM for each layer, and model all layers simultaneously with same community structure, but allow relationships between groups to vary.
   De Bacco Power Larremore Moore. Phys. Rev. E 95, 1981–10 (2017).

1 is preferred if nodes appear/disappear over time.
2,3 are preferred to solve the *layer interdependence problem*

Layer interdependence

more layers = better performance
(layer structure generated by same social mech.)

Are layers structurally similar? Complementary? Neither?

"Learn" a SBM from *m* layers; try to predict links of *m+1*.

12 layer social support network across 2 villages in South India.

Caterina De Bacco, Eleanor A. Power, Daniel B. Larremore, and Cristopher Moore.
"Community detection, link prediction, and layer interdependence in multilayer networks" Physical Review E 95 042317

Layer interdependence - malaria

more layers = worse performance
(layer structure generated by different biol. mech.)

can’t predict one HVR in the immune evasion genes by using others.

Caterina De Bacco, Eleanor A. Power, Daniel B. Larremore, and Cristopher Moore.
"Community detection, link prediction, and layer interdependence in multilayer networks” Physical Review E 95 042317

Advanced topic 4: metadata+communities

What is/are metadata?

How well do metadata explain the network structure? “BESTest”

Peel*, Larremore*, Clauset. Science Advances 3(5) e1602548. (2017) http://advances.sciencemag.org/content/3/5/e1602548.full

How do metadata relate to network structure? “neoSBM”

Peel*, Larremore*, Clauset. Science Advances 3(5) e1602548. (2017) http://advances.sciencemag.org/content/3/5/e1602548.full

Can we use metadata to guide community detection? “metadata assisted SBM”

Newman, Clauset. *Nature communications* 7 (2016). https://www.nature.com/ncomms/2016/160616/ncomms11863/full/ncomms11863.html

Can we find patterns in the metadata itself (themselves ugh)? Apply multilayer SBM

Peixoto. Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. Phys. Rev. X 4, 011047 (2014).

Blockmodel entropy significance test

How well do the metadata explain the network?

randomly assigned metadata
→ model gives no explanation, high H

metadata correlated with communities
→ model gives good explanation, low H

1. Divide the network G into groups according to metadata labels M.

2. Fit the maximum likelihood parameters of an *a posteriori* SBM and compute the entropy $H(G,M)$ of the corresponding ensemble.

3. Compare the entropy of this SBM ensemble to distribution of entropies from SBMs partitioned using <u>shuffled</u> metadata $\underline{M}$.

$$\text{p-value} = \Pr[H(G,\{\underline{M}\})) \leq H(G,M)]$$

Multiple network layers; multiple metadata attributes

Network	Status	Gender	Office	Practice	Law School
Friendship	$< 10^{-6}$	0.034	$< 10^{-6}$	0.033	0.134
Cowork	$< 10^{-3}$	0.094	$< 10^{-6}$	$< 10^{-6}$	0.922
Advice	$< 10^{-6}$	0.010	$< 10^{-6}$	$< 10^{-6}$	0.205

model = SBM

Multiple sets of metadata **significantly explain** multiple networks.
[Should one particular set of metadata be ground truth?]

https://piratepeel.github.io/code.html

BESTest accommodates many models of group structure

Network	Model SBM	Model DCSBM
Malaria 1	0.566	0.066
Malaria 2	0.064	0.126
Malaria 3	0.536	0.415
Malaria 4	0.588	0.570
Malaria 5	0.382	0.097
Malaria 6	0.275	0.817
Malaria 7	0.020	0.437
Malaria 8	0.464	0.143
Malaria 9	0.115	0.104

metadata = parasite origin

A negative result: parasite origin is irrelevant to genetic substring-sharing.

Malaria parasites *do not* have a strong strain structure, with implications for **diversifying selection** among parasites.

https://piratepeel.github.io/code.html

DBL, Clauset, A. & Buckee, C. O. A Network Approach to Analyzing Highly Recombinant Malaria Parasite Genes. PLoS Comp Bio 9, e1003268 (2013).

neoSBM

Choose between the SBM partition and the metadata partition.

$$\underbrace{\mathcal{L}_{\text{neoSBM}}}_{\substack{\text{neoSBM}\\ \text{log likelihood}}} = \underbrace{\mathcal{L}_{\text{SBM}}}_{\substack{\text{SBM}\\ \text{log likelihood}}} + \underbrace{f(\theta)}_{\text{cost}}$$

Log likelihood with parameterized prior:

θ is the parameter of a Bernoulli prior on whether the node is free to choose its own community or held fixed at its metadata label.

As θ increases, the cost of freeing a node decreases.

Varying θ in the unit interval **explores the space of partitions** between M and C.

https://piratepeel.github.io/code.html

Plant two different *kinds* of structure in a network

SBM with 8 groups and
two interesting 4-group partitions:

i. core-periphery

ii. assortative

https://piratepeel.github.io/code.html

The neoSBM identifies four interesting partitions

metadata SBM

https://piratepeel.github.io/code.html

The prior parameter changes the likelihood surface

Metadata-aware SBM

high school social network 🏫

795 students from an American high school + its feeder middle school

- $\mathbf{x} = \{$**grade 7-12**, ethnicity, gender$\}$

Newman & Clauset, *Nat. Comms.* 7, 11863 (2016)
Add Health network data, designed by Udry, Bearman & Harris
http://www.santafe.edu/~aaronc/

Metadata-aware SBM

high school social network

795 students from an American high school + its feeder middle school

- $\mathbf{x} = \{$**grade 7-12**, ethnicity, gender$\}$
- method finds a good partition between high-school and middle-school

$$\mathrm{NMI} = 0.881$$

- without metadata:

$$\mathrm{NMI} \in [0.105, 0.384]$$

Newman & Clauset, *Nat. Comms.* 7, 11863 (2016)
Add Health network data, designed by Udry, Bearman & Harris
http://www.santafe.edu/~aaronc/

Metadata-aware SBM

high school social network

795 students from an American high school + its feeder middle school

- $\mathbf{x} =$ {grade 7-12, **ethnicity**, gender}
- method finds a good partition between blacks and whites (with others scattered among)

  $$\mathrm{NMI} = 0.820$$

- without metadata:

  $$\mathrm{NMI} \in [0.120, 0.239]$$

Newman & Clauset, *Nat. Comms.* 7, 11863 (2016)
Add Health network data, designed by Udry, Bearman & Harris
http://www.santafe.edu/~aaronc/

Metadata-aware SBM

high school social network

795 students from an American high school + its feeder middle school

- $\mathbf{x} = \{$grade 7-12, ethnicity, **gender**$\}$
- method finds no good partition between males/females.
  instead, chooses a mixture of grade/ethnicity partitions

$$\mathrm{NMI} = 0.003$$

- without metadata:

$$\mathrm{NMI} \in [0.000, 0.010]$$

Newman & Clauset, *Nat. Comms.* 7, 11863 (2016)
Add Health network data, designed by Udry, Bearman & Harris
http://www.santafe.edu/~aaronc/

Task 4: ICON

The Colorado Index of Complex Networks (ICON)

ICON is a comprehensive index of research-quality network data sets from all domains of network science, including social, web, information, biological, ecological, connectome, transportation, and technological networks.

Each network record in the index is annotated with and searchable or browsable by its graph properties, description, size, etc., and many records include links to multiple networks. The contents of ICON are curated by volunteer experts from Prof. Aaron Clauset's research group at the University of Colorado Boulder.

Click on the NETWORKS tab above to get started.

Entries found: 606 Networks found: 4416

http://icon.colorado.edu/

Other things to know about 1: “The Louvain Method”

If your network is *really* big. (Millions of nodes, Billions of edges)

Take ClausetNewmanMoore’s approach for greedy Q maximization and find small groups. Run the code again on those groups… And again…

Advantage: fast! big! 🚀

Disadvantage: inherits the assumptions of modularity. (clustering vs modeling)

6K cites. People like it!

https://en.wikipedia.org/wiki/Louvain_Modularity https://arxiv.org/pdf/0803.0476.pdf

Other things to know about 2: InfoMap

Imagine a random walker on a network.

A description of her walk can be compressed if the network has regions in which the random walker tends to stay for a long time.

Minimizing the “map equation” over all possible network partitions is the same as finding the best codebook.

http://www.mapequation.org/apps/MapDemo.html

http://www.mapequation.org/code.html

Outlook

Simply put, we have amazingly powerful tools that did not exist 15 years ago.

Many are principled, statistically rigorous, and we learn more all the time. Those that aren't statistically rigorous are really, really fast.

There is no multiple regression for networks.

"Controlling for C, how important is X in predicting Y?"

Tradeoffs between general and bespoke methods are still being explored.

Outside of SBM, Modularity, Louvain, Infomap, it's a wild west.

Methodologists are keen to be challenged by new problem types.

[Elizabeth Bruch nerd sniped me!] https://xkcd.com/356/

THERE'S A CERTAIN TYPE OF BRAIN THAT'S EASILY DISABLED.
IF YOU SHOW IT AN INTERESTING PROBLEM, IT INVOLUNTARILY DROPS EVERYTHING ELSE TO WORK ON IT.
THIS HAS LED ME TO INVENT A NEW SPORT: NERD SNIPING.
SEE THAT PHYSICIST CROSSING THE ROAD?
HEY!
On this infinite grid of ideal one-ohm resistors,
what's the equivalent resistance between the two marked nodes?
IT'S... HMM. INTERESTING. MAYBE IF YOU START WITH ... NO, WAIT. HMM...YOU COULD—
FOOOOM
I WILL HAVE NO PART IN THIS.
C'MON, MAKE A SIGN. IT'S FUN! PHYSICISTS ARE TWO POINTS, MATHEMATICIANS THREE.

Many uses for the same techniques. cf regression

Treat the network like a system:

Extrapolation. Make predictions for as-yet unseen nodes (in “space” or time).
Interpolation. Identify missing links.
Generalization. Nodes of this type are like others of the same type.

Treat the network like an artifact:

Mechanisms. How did this network arise? What rules governed its assembly?
Explanations. Coarse-graining or compression.

Treat the network like a means to an end; an intermediate data structure:

Useful division. Need groups so that we can assign treatments in an A/B test.
Simplification. Downstream regression model needs ranks or groups.

https://www.youtube.com/watch?v=1v5_CzdRVAc

Rankings and linear hierarchies

Many uses for the same techniques. cf regression

Treat the network like a system:

Extrapolation. Make predictions for as-yet unseen nodes (in “space” or time).
Interpolation. Identify missing links.
Generalization. Nodes of this type are like others of the same type.

Treat the network like an artifact:

Mechanisms. How did this network arise? What rules governed its assembly?
Explanations. Coarse-graining or compression.

Treat the network like a means to an end; an intermediate data structure:

Useful division. Need groups so that we can assign treatments in an A/B test.
Simplification. Downstream regression model needs ranks or groups.

The idea of rankings—pervasive!

Assumptions:

1. Competitors have some intrinsic quality (or vector of qualities).
2. Interactions can (stochastically) reveal differences in qualities.
3. Competitions are pair-wise. (Lee Sedol vs. AlphaGo; Astros vs. Dodgers)

In other words: outcomes are generated by a stochastic process, which is some function of the positions of the competitors.

Systems of dominance

social

physical

mental

financial

Systems of endorsement

Assumptions:

1. Endorsers have some intrinsic quality.
2. Interactions can reveal differences in qualities.
3. Endorsements are pair-wise.

Clauset, Arbesman, Larremore. *Science Advances* 1, e1400005 (2015).

Systems of endorsement

Latent position can be revealed by **dominance** or **endorsement** interactions.

Figure: Larremore, Hébert-Dufresne, Power. *Draft.*

Data: Power. *Nature Human Behaviour* 1, 0057 (2017)

Win-Loss is not satisfactory: schedule matters

Beating the grandmaster counts for more than beating a novice.

Win and loss tallies don't take this "schedule difficulty" into account. Put differently, win-loss records **leave information on the table.**

One way to make use of this information:

i* beats *j implies $s_i > s_j$

Therefore if we have a whole list of outcomes, we can try to find a total ordering that breaks as few of these implications as possible.

A_{ij} = number of times that *i* beat *j*.

minimum violation ranking: sort *A*.

MVR: non-unique & rough optimization landscape

There is no guarantee of a unique minimizing ranking *s*.

Space of ordinal rankings has n! elements—usually use MCMC to search.

Slow.

Ordinal. No ties. No interpretability of rank differences.

What are other premises on which
we can base a ranking model?

minimum violation ranking: sort *A*.

Embeddings vs Orderings

Ordering is ordering. 1, 2, 3, …

Embedding is assigning a position to each node. 1, 1.2, 7, 20, 21, 21.2, …

Which one is right for me?

Depends on the use. Consider that you can always go from an embedding to an ordering, if you have a rule for breaking ties.

Embeddings and Orderings: MVR & Agony

What if you allowed for ties and then
ran Min. Viol. Ranking? What would happen?

MVR: uniform cost (1 per edge).
Agony: generic cost function.

Embeddings and Orderings 1: PageRank

PageRank defines scalar rank recursively:

important pages are those that are linked to by important pages.

- Great at finding the top 3 but low interpretability of the PageRank scores.

The PageRank Citation Ranking: Bringing Order to the Web

January 29, 1998

Abstract

The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them.

We compare PageRank to an idealized random Web surfer. We show how to efficiently compute PageRank for large numbers of pages. And, we show how to apply PageRank to search and to user navigation.

The Anatomy of a Large-Scale Hypertextual Web Search Engine

Sergey Brin and Lawrence Page

Computer Science Department,
Stanford University, Stanford, CA 94305, USA
sergey@cs.stanford.edu and page@cs.stanford.edu

Abstract

In this paper, we present Google, a prototype of a large-scale search engine which makes heavy use of the structure present in hypertext. Google is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems. The prototype with a full

Eigenvector centrality on [A + very weak all-to-all links].

Embeddings and Orderings 2: Ball & Newman

Ball & Newman inferred ordinal rank via random graph + ranks model:

infer parameters of people's attachment preferences & ranks.

- Identified the need to learn from reciprocated friendships.
- Found that in AdHealth data, teens link to others of *nearby* social status.

Generative model + EM algorithm to fit it to data.
Integer ordering of nodes

Ball, Newman. Network Science 1, 16-30 (2013)

Embeddings and Orderings 3: Niche Models

Niche Models embed species in a latent space based on feeding preferences:

most species feed from narrow range in a 1-dim. space (~body size).

- Great for food webs. Inference models v slow for all but small networks.

Figure 1 Diagram of the niche model. Each of ***S*** species (for example, ***S*** = 6, each shown as an inverted triangle) is assigned a 'niche value' parameter (n_i) drawn uniformly from the interval [0,1]. Species ***i*** consumes all species falling in a range (r_i) that is placed by uniformly drawing the centre of the range (c_i) from [r_i/2, n_i]. This permits looping and cannibalism by allowing up to half of r_i to include values $\geqslant n_i$. The size of r_i is assigned by using a beta function to randomly draw values from [0,1] whose expected value is 2***C*** and then multiplying that value by n_i [expected E(n_i) = 0.5] to obtain the desired ***C***. A beta distribution with $\alpha = 1$ has the form $f(x|1, \beta) = \beta(1-x)^{\beta-1}$, $0 < x < 1$, 0 otherwise, and $E(X) = 1/(1+\beta)$. In this case, $x = 1-(1-y)^{1/\beta}$ is a random variable from the beta distribution if y is a uniform random variable and β is chosen to obtain the desired expected value. We chose this form because of its simplicity and ease of calculation. The fundamental generality of species i is measured by r_i. The number of species falling within r_i measures realized generality. Occasionally, model-generated webs contain completely disconnected species or trophically identical species. Such species are eliminated and replaced until the web is free of such species. The species with the smallest n_i has $r_i = 0$ so that every web has at least one basal species.

Williams & Martinez. Nature 404.6774 (2000).

Embeddings and Orderings 4: BTL

Bradley-Terry-Luce embed players in a 1D space. Outcome direction is simply:

$$P(i \to j) = \frac{\gamma_i}{\gamma_i + \gamma_j}$$

- Provably avoids <u>non-transitive</u> properties
- Great when lots of data per interaction.

Honestly? Best to google “BTL model.”

Supplement: Non-transitive dice

- 3 (or more) dice {A,B,C}
- faces chosen so that they have the property:
  - A>B more than half the time.
  - B>C more than half the time.
  - C>A more than half the time (?!)

https://en.wikipedia.org/wiki/Nontransitive_dice

A great gift for your favorite nerd's desk!
Go to the makerspace and laserbeam your own!

Embeddings and Orderings 4: BTL

Bradley-Terry-Luce embed players in a 1D space. Outcome direction is simply:

$$P(i \to j) = \frac{\gamma_i}{\gamma_i + \gamma_j}$$

- Provably avoids <u>non-transitive</u> properties
- Great when lots of data per interaction.

Pairwise ranking is really nice for ordering large sets of preferences too, and this model specifically models the probability that the preference will be for *i* over *j*.

Iterative algorithms exist. Needs a little regularization so the winningest winners don't fly off to infinity.

A few more minutes on:
SpringRank—fast, interpretable, predictive.

Honestly? Best to google "BTL model."

SpringRank

Each directed edge = directed spring

Relax and let the springs decide the ranks

$$H(s) = \frac{1}{2} \sum_{i,j=1}^{N} \mu_{ij} A_{ij} \left(s_i - s_j - \ell\right)^2$$

SpringRank Hamiltonian: energy of the system, given the positions *s.*

Because the springs are linear, the potential is quadratic.

The SR Hamiltonian is *convex* in s.

$$\nabla H(s) = 0$$

unique.

rural
social
support
parakeets
CS hiring
1
0.5
0
-0.5
-1
-1.5
-2
0
10
20
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
0
2
4
1
0.5
0
-0.5
0
10

Things we talked about at the end in Q&A

Satellites used to predict measles fluctuations:
Bharti, Nita, et al. "Explaining seasonal fluctuations of measles in Niger using nighttime lights imagery." Science 334.6061 (2011): 1424-1427.

Change-point detection in a sequence of networks
https://piratepeel.github.io/proximitynetwork.html

Many uses for the same techniques. cf regression

Treat the network like a system:

Extrapolation. Make predictions for as-yet unseen nodes (in “space” or time).
Interpolation. Identify missing links.
Generalization. Nodes of this type are like others of the same type.

Treat the network like an artifact:

Mechanisms. How did this network arise? What rules governed its assembly?
Explanations. Coarse-graining or compression.

Treat the network like a means to an end; an intermediate data structure:

Useful division. Need groups so that we can assign treatments in an A/B test.
Simplification. Downstream regression model needs ranks or groups.

https://www.youtube.com/watch?v=1v5_CzdRVAc

Thank you

@danlarremore
daniel.larremore@colorado.edu