Model selection in networks

Matija Piškorec[1]

[1]Ruđer Bošković Institute, Zagreb, Croatia

Qualification doctoral exam on Faculty of Electrical Engineering and Computing, November 23, 2015

Part I

Complex networks

What is a network?

Network

A set of *nodes* and a set of *connections* between them, along with any number of their *properties*.

Why are networks useful?

Many real world systems can be represented as networks.

Figure: Social networks.

Why are networks useful?

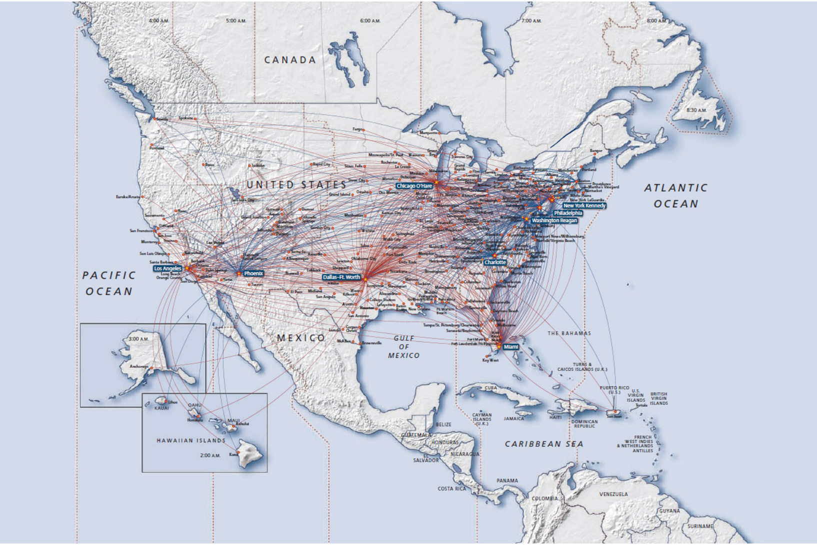


Figure: Airline network.

Why are networks useful?

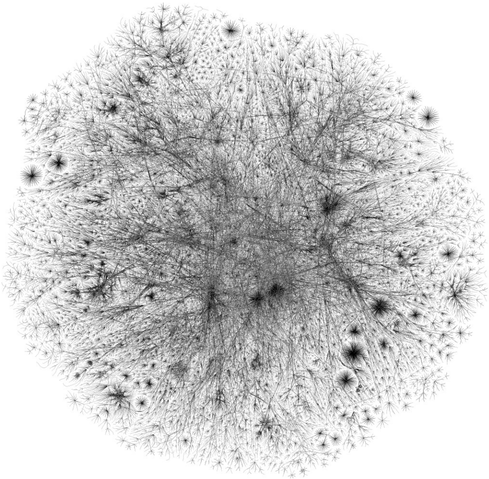


Figure: Internet.

Why are networks useful?

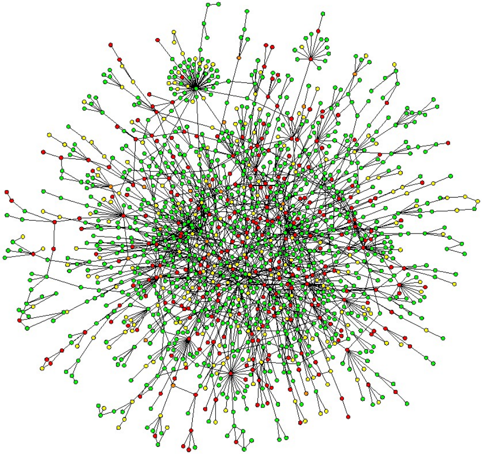


Figure: Yeast protein interaction network.

Universal properties of real networks

Heavy tail degree distribution

Number of connections k that each node has is distributed according to a *power-law* with $P_k \sim k^{-\gamma}$ as $k \to \infty$

Small diameter

Maximum length of the path that connects any two nodes in the network scales logarithmically with the number of nodes.

Clustering of nodes

Clustering manifests on a microscale where nodes form triangles, on a mesoscale where they form communities, and on a global scale where they form core-periphery structure of the network.

Part II

Model selection

Models on networks

- Growth models [1] (for example, *preferential attachment* model)
- Community models [2]
- Contact processes
  - Epidemic spreading (for example, *Susceptible-Infected-Recovered* model)
  - Social influence [3]
  - Information diffusion [4]

[1] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, 1999.

[2] M. Girvan and M. Newman, "Community structure in social and biological networks," *PNAS*, 2002.

[3] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and correlation in social networks," 2008

[4] Y. Moreno, M. Nekovee, and A. Pacheco, "Dynamics of rumor spreading in complex networks," *Physical Review E*, 2004.

Model selection

The central goal of model selection is to select a model or a class of models which generalizes best to unobserved data. In general this is not easy because of the *bias/variance tradeoff*.

The crucial factor that determines the model's generalization performance is its *complexity*, which is a measure of model's degrees of freedom which allow it to fit many different datasets.

Model selection and parameter estimation

Result of parameter estimation is always *one* specific hypothesis, while model selection gives us a *set* of hypothesis. For example, a set of hypotheses might be a set of all k^{th} degree polynomials, or, even more generally, a set of all polynomials.

Some benefits of performing model selection instead of parameter estimation:

- **Finding a "general theory".** For example, when we want to select a model that performs well under a variety of circumstances for which particular parameters of a model may differ.
- **Easier creation of model hierarchy.** If our model space has a hierarchical structure we can easily infer hierarchical models by evaluating subsets of models and composing them into hierarchy.

Bayesian model selection

If we consider a set of models $\{\mathcal{M}_i\}$, where each model $\mathcal{M}_i$ is defined as a probability distribution over an observed data set $\mathcal{D}$, then we are searching for a model with the largest posterior distribution:

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i) \tag{1}$$

The prior $p(\mathcal{M}_i)$ expresses a prior probability of different models before any data is observed. *Model evidence* (also known as *marginal likelihood*) $p(\mathcal{D}|\mathcal{M}_i)$ expresses the preference for different models given the data:

$$p(\mathcal{D}|\mathcal{M}_i) = \int \underbrace{p(\mathcal{D}|\theta, \mathcal{M}_i)}_{\textit{likelihood}} \underbrace{p(\theta|\mathcal{M}_i)}_{\textit{prior}} \, d\theta \tag{2}$$

Note that the parameters θ of a model are marginalized (integrated) out. This has three consequences (two positive and one negative)!

Model evidence

Consequences of the marginalization of the parameters θ from the model evidence:

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\theta, \mathcal{M}_i) p(\theta|\mathcal{M}_i) d\theta$$

1) this allows to meaningfully compare models irrespective of their parameterizations
2) this implicitly restricts the complexity of the model, and in this way prevents the overfitting of the model to the data
3) marginalization of the parameters is computationally very hard problem

Model selection criteria - maximum likelihood

Complete marginalization of model evidence is often computationally infeasible. Common approach is to just find the parameters for which likelihood function achieves maximum.

Better approach is to take into account both the maximum value of likelihood function (*fitness*) and additional term which accounts for the shape of the likelihood function (*complexity*).

Model selection criteria

Following model selection criteria use the maximum value of likelihood function in addition to different complexity terms in order to approximate model evidence:

$$\begin{aligned}\text{Akaike Information Criterion} &= -2\ln f(y|\hat{\theta}) + 2k \\ \text{Bayesian Information Criterion} &= -2\ln f(y|\hat{\theta}) + k\ln N \\ \text{Rissanen's Stochastic Complexity} &= \underbrace{-\ln f(y|\hat{\theta})}_{\textit{fitness}} + \underbrace{\frac{k}{2}\ln N}_{\textit{complexity}}\end{aligned} \tag{3}$$

Where k is the number of parameters of the model and N is the number of observations. Here $f(y|\hat{\theta})$ corresponds to likelihood $p(\mathcal{D}|\theta, \mathcal{M}_i)$ in equation 2.

However, AIC, BIC and RSC do not consider the functional form of the model, which can be different even for models with the same number of parameters!

Model selection criteria - examples for AIC and BIC

We will illustrate the influence of functional form of the model with the following examples from psychophysics [5]:

$$
\begin{aligned}
y &= ax^b + \text{error (Stevens' model)} \\
y &= a\ln(x+b) + \text{error (Fechner's model)}
\end{aligned}
\tag{4}
$$

Model fitted	Data from Stevens	Data from Fencher
Stevens	100%	63%
Fencher	0%	37%

Table: Model selection for two models using AIC and BIC. Although both models have the same number of parameters, Steven's model fits data more easily, even in artificial case when data is generated from Fencher's model!

[5] J. Myung, V. Balasubramanian and A. Pitt, "Counting probability distributions: Differential geometry and model selection," *PNAS*, 2000.

Model selection criteria

Minimum Description Length [6] accounts for the model's functional form through Fisher Information Matrix $I_{ij}(\theta)$:

$$\mathrm{MDL} = \underbrace{-\ln f(y|\hat{\theta})}_{\text{fitness}} + \underbrace{\frac{k}{2}\ln\left(\frac{N}{2\pi}\right) + \ln\int d\theta\sqrt{\det I(\theta)}}_{\text{complexity of a model family } f} \tag{5}$$

Model fitted	Data from Stevens	Data from Fencher
Stevens	99%	2%
Fencher	1%	98%

Table: Model selection for two models using MDL.

MDL is the length in bits of the shortest possible code which describes the data generated by a model lying within the family f

[6] P. Grunwald, "Model selection based on minimum description length," *Journal of Mathematical Psychology*, 2000.

Other model selection criteria

Structural risk minimization [7] Where complexity of a model is measured by *Vapnik-Chervonenkis dimension*.

False Discovery Rate [8] Used in multiple hypothesis testing. It works by controlling the expected proportion of rejected null-hypotheses which were in fact correct ("false discoveries").

Cross-validation [9] Where a model is repeatedly learned and tested on two disjoint subsets of the observed data, in hope to select for a model that will have good predictive accuracy on unobserved data.

[7] V. N. Vapnik, *Statistical Learning Theory*, 1989.

[8] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerfull approach to multiple testing," *Journal of the Royal Statistical Society*, 1995.

[9] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society*, 1974.

Part III

Inference in networks

Inference of network structure

Network growth models. Inference using maximum likelihood approach through efficient Monte Carlo sampling [10] [11].

Community structure models. Two representations which allow definition of hierarchical models and for which efficient inference methods were developed are *Kronecker graphs* [12] and *block models* [13].

[10] I. Bezáková, A. Kalai, and R. Santhanam, "Graph model selection using maximum likelihood," 2006.

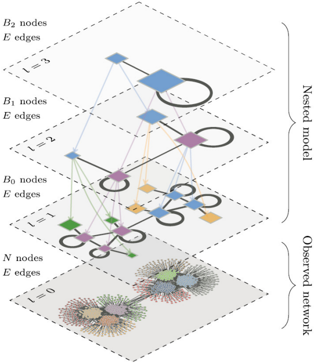
[11] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, "Microscopic evolution of social networks," 2008.

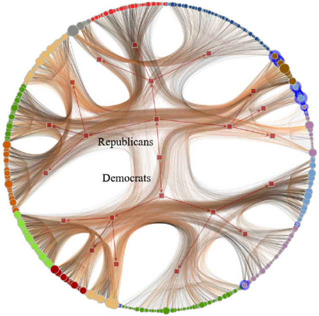
[12] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, "Kronecker graphs: An approach to modeling networks," *Journal of Machine Learning Research*, 2010.

[13] T. Peixoto, "Parsimonious module inference in large networks," *Physical Review Letters*, 2013.

Inference of network structure - block models

Block models are convenient for representing hierarchical community structure, and they can be efficiently inferred using minimum description length [14] [15].





[14] T. Peixoto, "Parsimonious module inference in large networks," *Physical Review Letters*, 2013.

[15] T. Peixoto, "Hierarchical block structures and high-resolution model selection in large networks," *Physical Review X*, 2014.

Inference of network structure from information diffusion

Sometimes the information on network structure is lacking and the information on dynamics on network (such as information diffusion) is used to infer it. There are several algorithms which give a maximum likelihood spreading cascade, using different optimization strategies:

- *CoNNIe* and *NetRate* use convex programming [16] [17]
- *NetInf* uses submodular function optimization [18]
- *InfoPath* uses stochastic gradient descent [19]

[16] S. Myers and J. Leskovec, "On the convexity of latent social network inference," 2010.

[17] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," 2011.

[18] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," *ACM Transactions on Knowledge Discovery from Data*, 2012.

[19] M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf, "Structure and dynamics of information pathways in online media," 2013.

Inference of processes on networks

In comparison to inference of network structure, inference of processes on networks still lacks a suitable representation which would allow inference of a broad range of dynamical models using an unified probabilistic framework.

Inference of processes on networks

General network dynamics equation for processes on networks [20]:

$$\frac{dx_i}{dt} = M_0(x_i(t)) + \sum_{j=1}^{N} A_{ij} M_1(x_i(t)) M_2(x_j(t)) \tag{6}$$

This equation can describe epidemic processes, biochemical dynamics, birth-death processes and gene regulatory dynamics [21].

Minimal functional form of this equation can be infered using aggregated features of the *transient response* $x_i(t)$ and the *response matrix* G_{ij}, which describe a response of the system after perturbation [22].

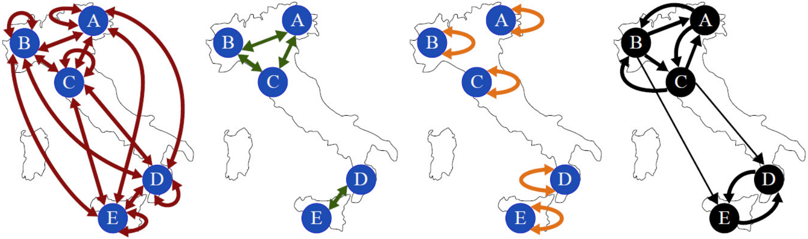
[20] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical processes on complex networks*, 2008.

[21] B. Barzel and A.-L. Barabási, “Universality in network dynamics,” *Nature Physics*, 2013.

[22] B. Barzel, Y.-Y. Liu, and A.-L. Barabási, “Constructing minimal models for complex system dynamics,” *Nature Communications*, 2015.

Inference of human trails on the Web

Inference of human trails on the Web using Bayesian inference. Hypotheses are specified as Markov chains [23] [24].



[23] P. Singer, D. Helic, B. Taraghi, and M. Strohmaier, "Detecting memory and structure in human navigation patterns using markov chain models of varying order," *PLoS ONE*, vol. 9, no. 7, 2014.

[24] P. Singer, D. Helic, A. Hotho, and M. Strohmaier, "Hyptrails: A bayesian approach for comparing hypotheses about human trails on the web," in *Proceedings of the 24th International Conference on World Wide Web*, 2015.

Challenges ahead

1. **Beyond maximum likelihood** Development of model selection procedures that rely on solid probabilistic foundations.
2. **Nonparametric approach** Emphasis on hierarchical models that can be inferred in nonparametric way.
3. **Dynamics rather than structure** Emphasis on processes on networks.
4. **Motivation** Data-driven and problem-driven approach.