Statistical Inference for Complex Networks

Santa Fe Institute
3-5 December 2008

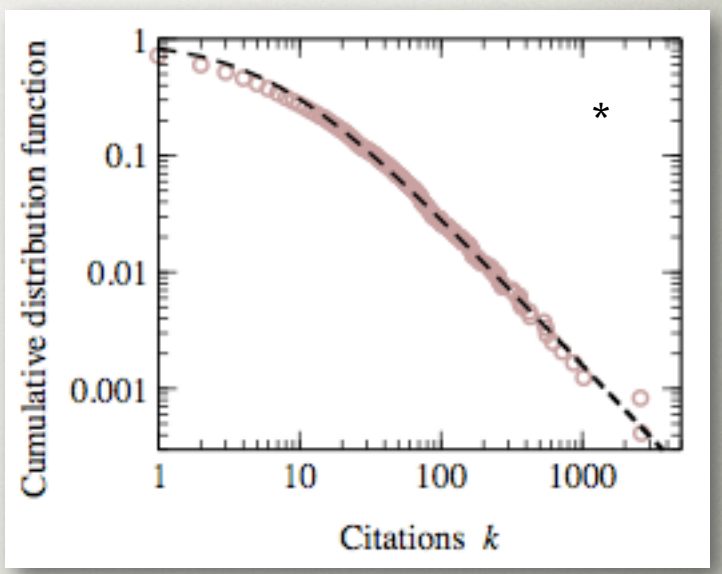
Simple Models

1. "traditional" scientific approach
    - de Solla Price's "cumulative advantage" model of citation network growth
2. small number of parameters
3. often analytically solvable
4. mechanistic explanation of data

Successes: an example

de Solla Price's "cumulative advantage" model

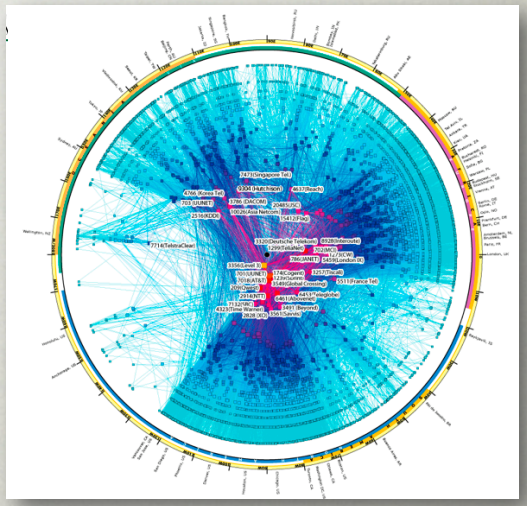
1. easy to understand
2. addresses cause & effect
3. predicts distributions of several quantities
4. agreement with data seems good



* from Newman (2008) arXiv:0809.0522

Failures: many

1. indirect comparisons with data
   - compare summary statistics of data with predictions of model
2. but for some patterns, many explanations
   - e.g., degree distributions (dozens of models)
   - new data often not possible
   - model selection & complexity control become hard
3. omit much real complexity
   - modules, hierarchies, spatial organization, etc.



Complex Models

1. common in machine learning, engineering
2. many! parameters
3. often not analytically solvable
4. often “data models” not “process models”
5. often in *likelihood* framework

Promises & Problems

1. can capture complex structures, go beyond "summary statistics" like degree distribution etc.
2. model fitting, testing, comparison all easier
3. direct comparison with network data
4. can avoid overfitting, e.g. by cross-validation and predicting out-of-sample data
5. but models often labor intensive & computationally expensive to fit, text, etc.
6. and how to interpret results for science? often no cause & effect

A Few Questions

1. What to do when can't put model into likelihood framework? (e.g., evolutionary models)
2. Model comparison: when is a fancier model worth it?
3. How to do "science", establish cause & effect, not just fit the data?
4. What are the BIG, cross-disciplinary questions, driven by new data?
    - heterogeneous structure: modules, hierarchy?
    - beyond topology: time, duration, type of contacts; flows and capacities of nodes and edges?
    - network evolution over time?