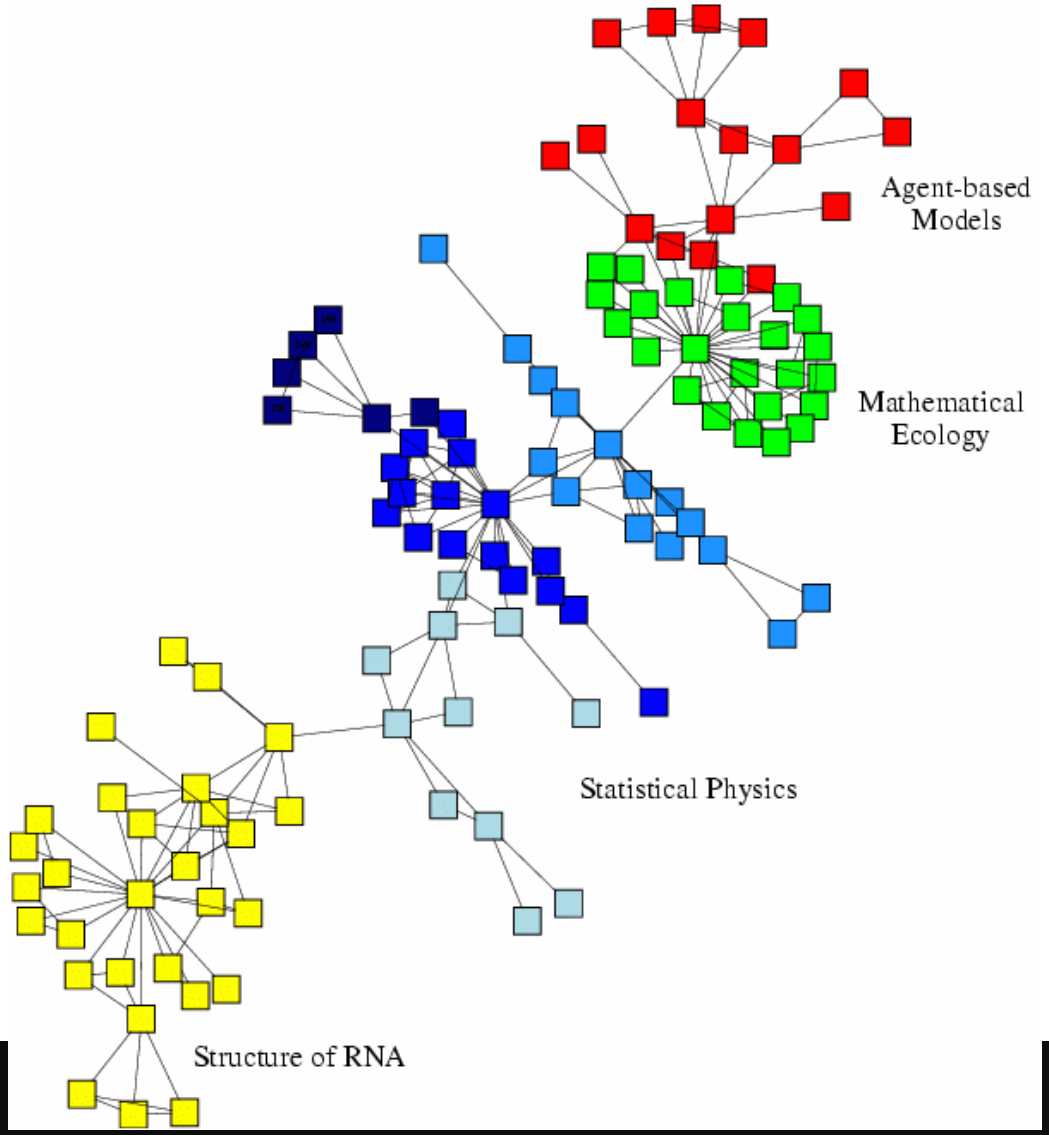
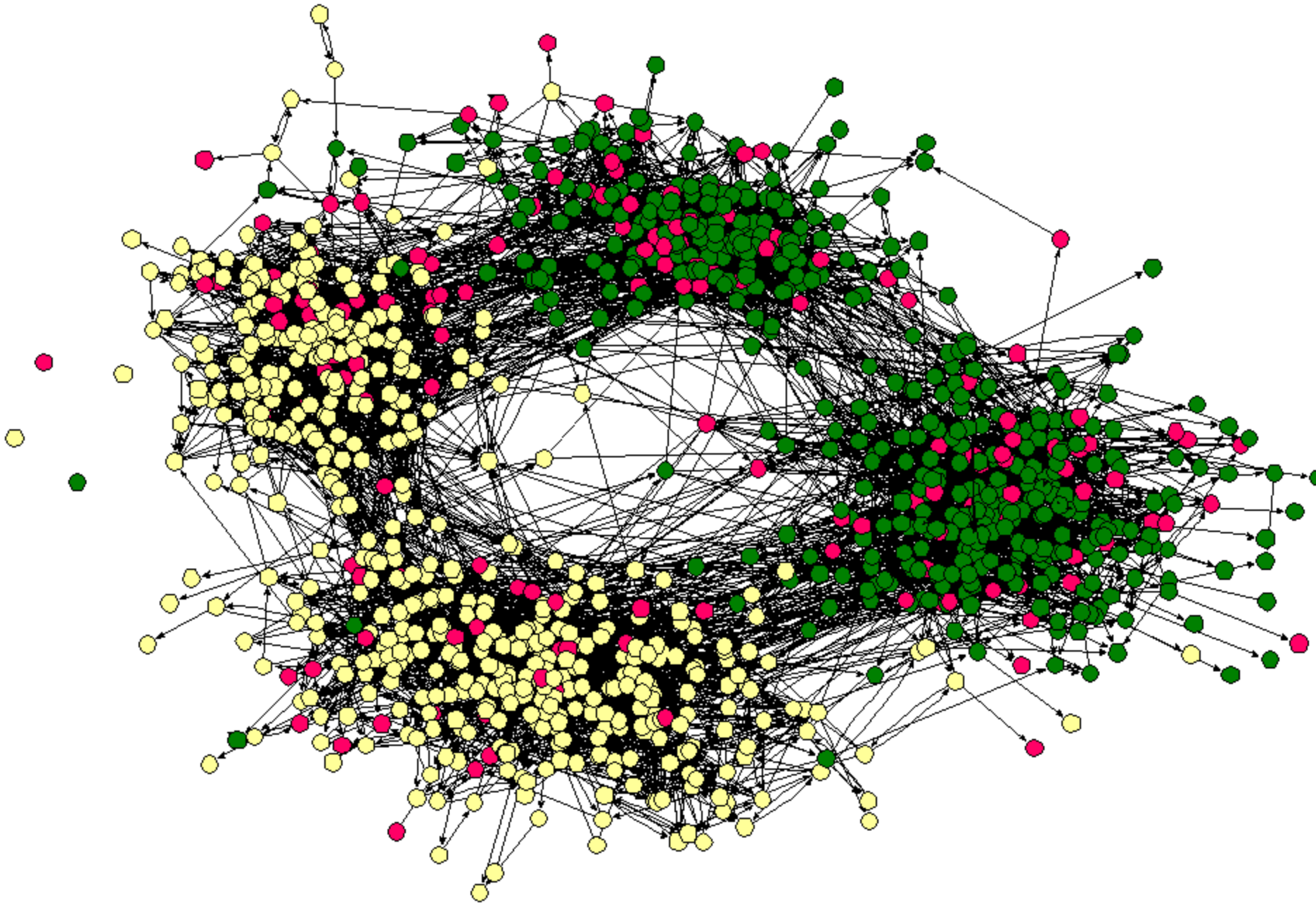
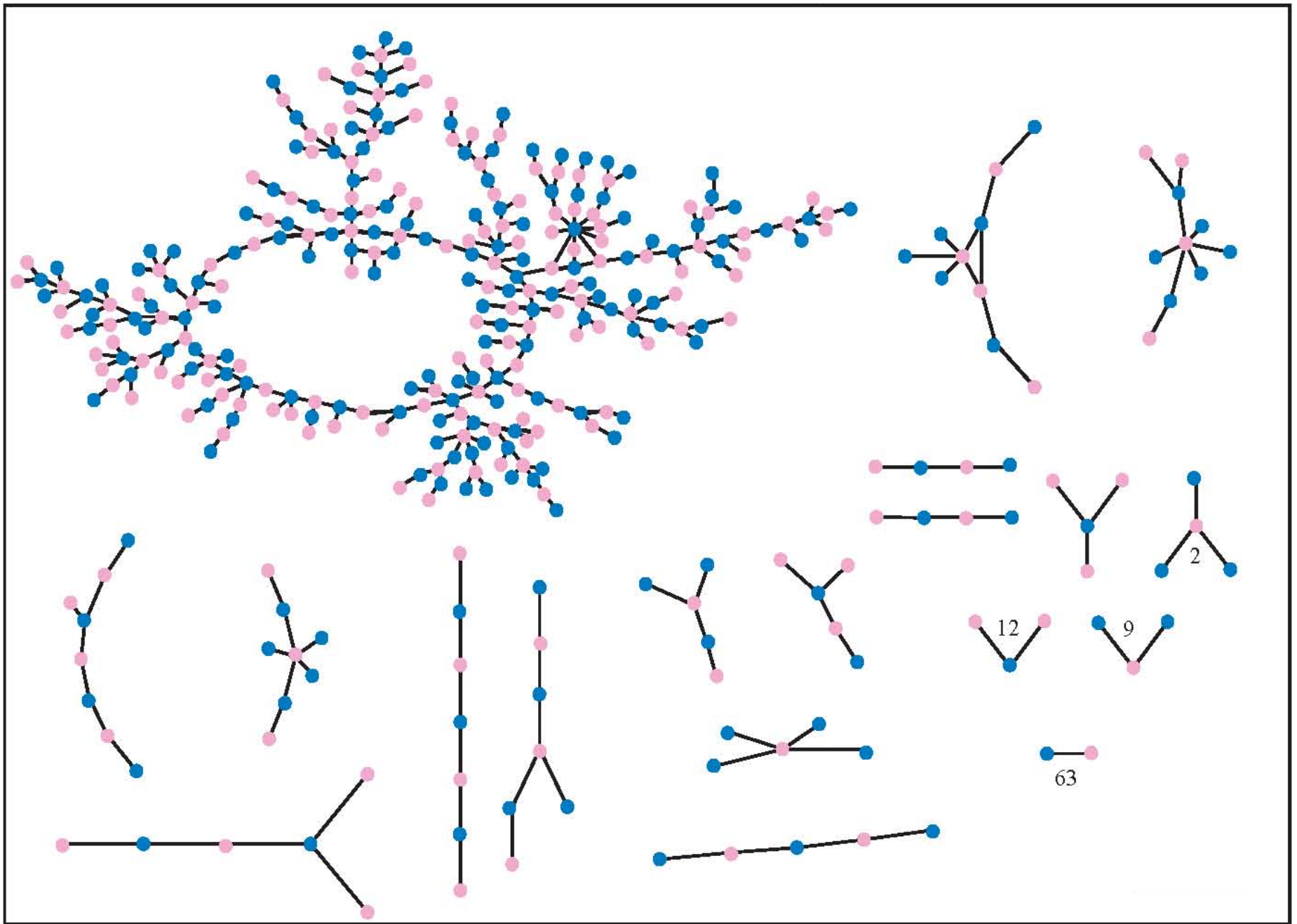
Foundations of Data Science

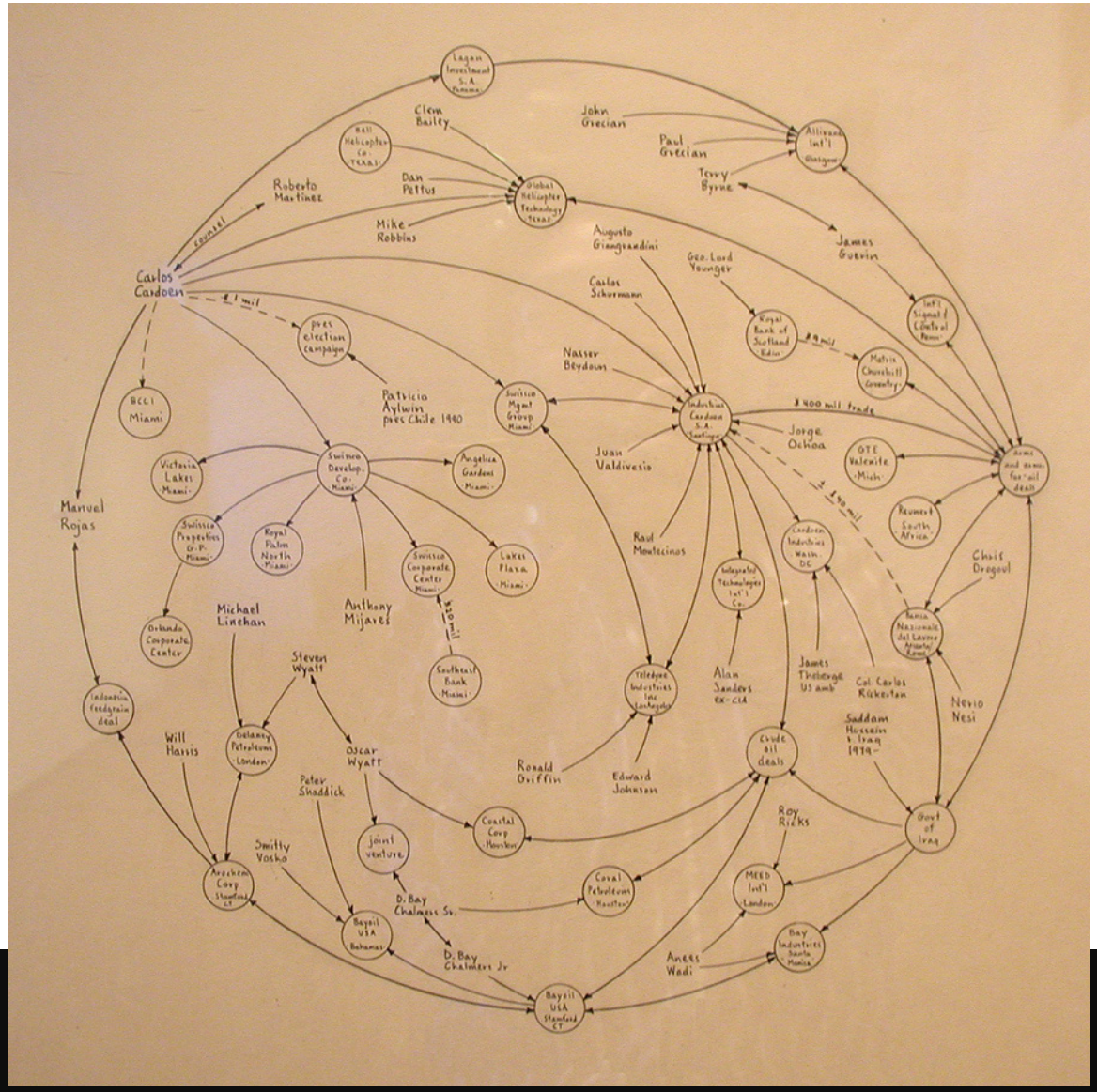
Rumi Chunara, PhD

CS3943/9223

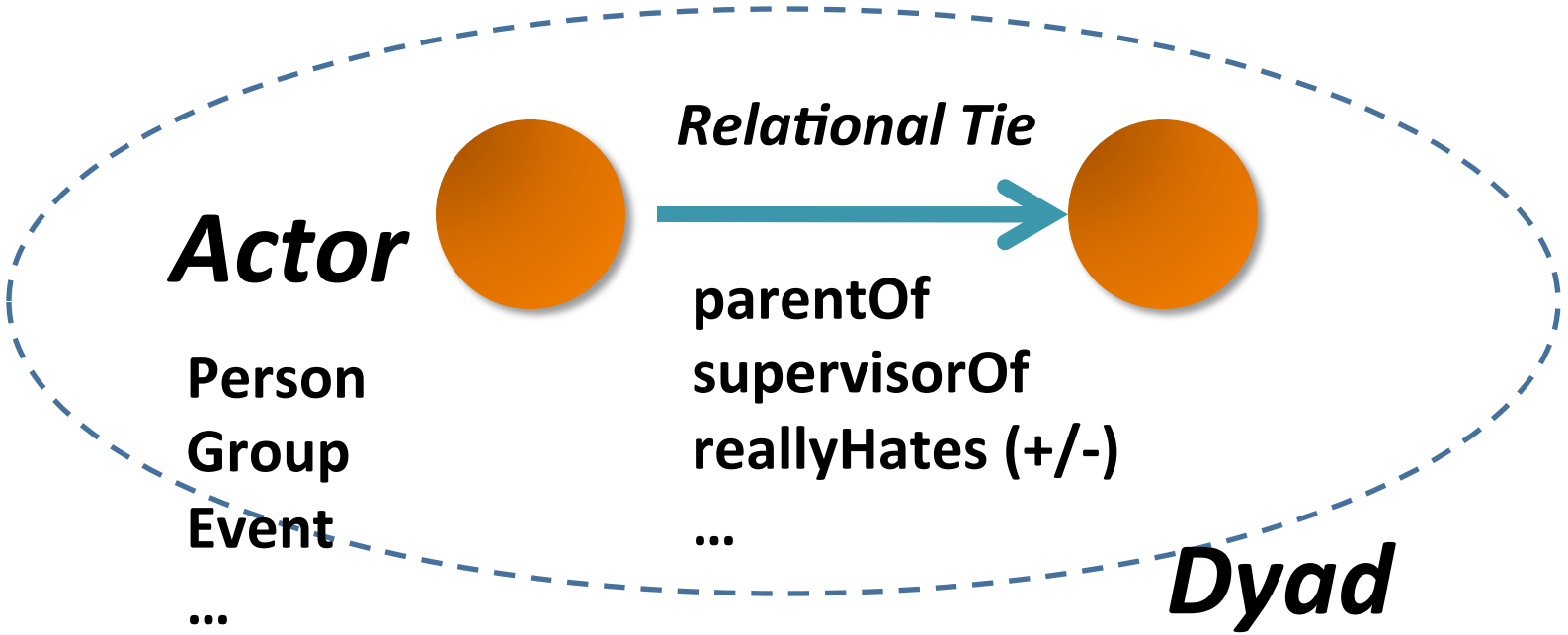

Agent-based
Models
Mathematical
Ecology
Statistical Physics
Structure of RNA




2
12
9
63


Lagan Investment S.A. Panama
Clem Bailey
Bell Helicopter Co. Texas
Dan Pettus
Mike Robbins
Global Helicopter Technology Texas
John Grecian
Paul Grecian
Allivan Int'l Glasgow
Terry Byrne
Roberto Martinez
counsel
Carlos Cardoen
$1 mil
pres election campaign
Patricio Aylwin pres Chile 1990
BCCI Miami
Augusto Giangrandini
Geo. Lord Younger
Carlos Schurmann
James Guerin
Int'l Signal & Control Penn.
Royal Bank of Scotland Edin.
$9 mil
Matrix Churchill Coventry
Nasser Beydoun
Swisco Mgmt Group Miami
Industrias Cardoen S.A. Santiago
$400 mil trade
Jorge Ochoa
Juan Valdivesia
OTE Valemite Mich.
Arms and ammo for mil deals
Victoria Lakes Miami
Swisco Develop. Co. Miami
Angelica Gardens Miami
Manuel Rojas
Swissco Properties G.P. Miami
Royal Palm North Miami
Swissco Corporate Center Miami
Lakes Plaza Miami
Raul Moutecinos
Cardoen Industries Wash. DC
Telegroup Technologies Int'l Co.
Reumert South Africa
Chris Dragoul
Banco Nacional del Ejercito Argentina Rome
Orlando Corporate Center
Michael Linehan
Anthony Mijares
$10 mil
Southeast Bank Miami
Steven Wyatt
Indonesia feedgrain deal
Will Harris
Delaney Petroleum London
Oscar Wyatt
Peter Shaddick
Telefono Industries Int. Santiago
Alan Sanders ex-CIA
James Theberge US amb
Col. Carlos Rickerton
Nevio Nesi
Saddam Hussein v. Iraq 1979-
Crude oil deals
Ronald Oriffin
Edward Johnson
Govt of Iraq
Roy Ricks
Coastal Corp Houston
joint venture
Smitty Vosko
Arochem Corp Stamford CT
Coral Petroleum Houston
MEED Int'l London
D. Bay Chalmers Sr.
Bayoil USA Bahamas
D. Bay Chalmers Jr
Anees Wadi
Bay Industries Santa Monica
Bayoil USA Stamford CT

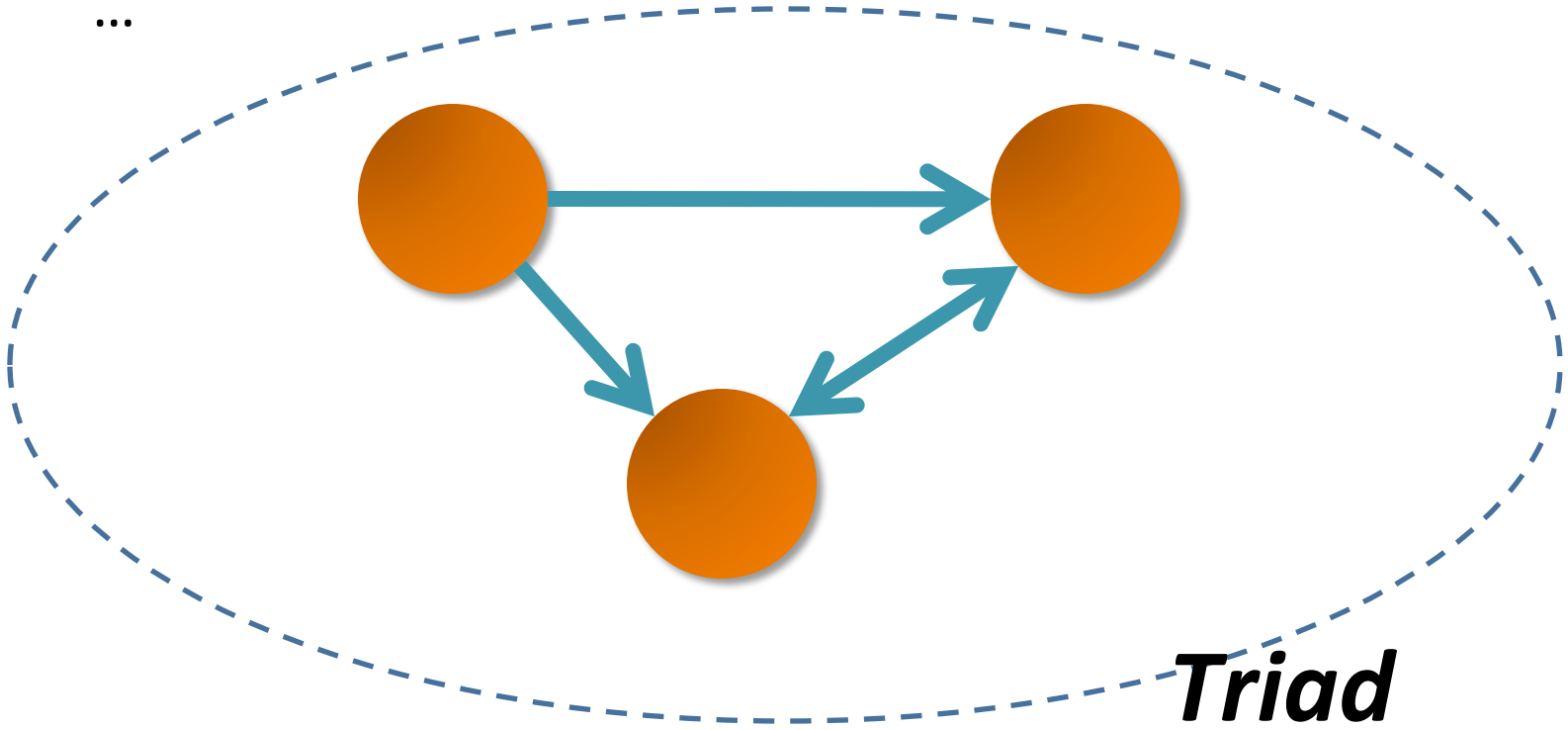
Vocabulary Lesson



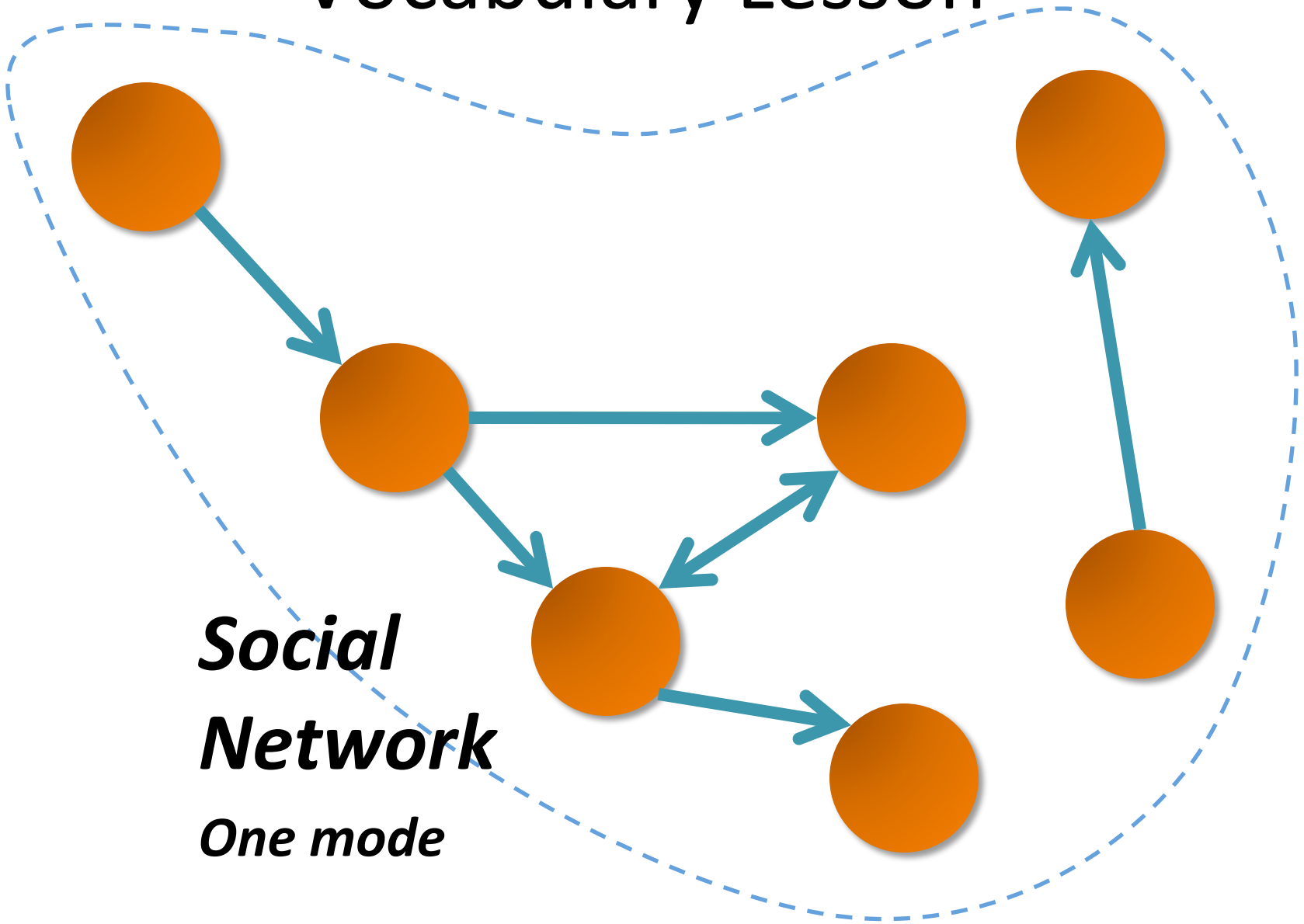
***Relation*: collection of ties of a specific type (every parentOf tie)**

Vocabulary Lesson

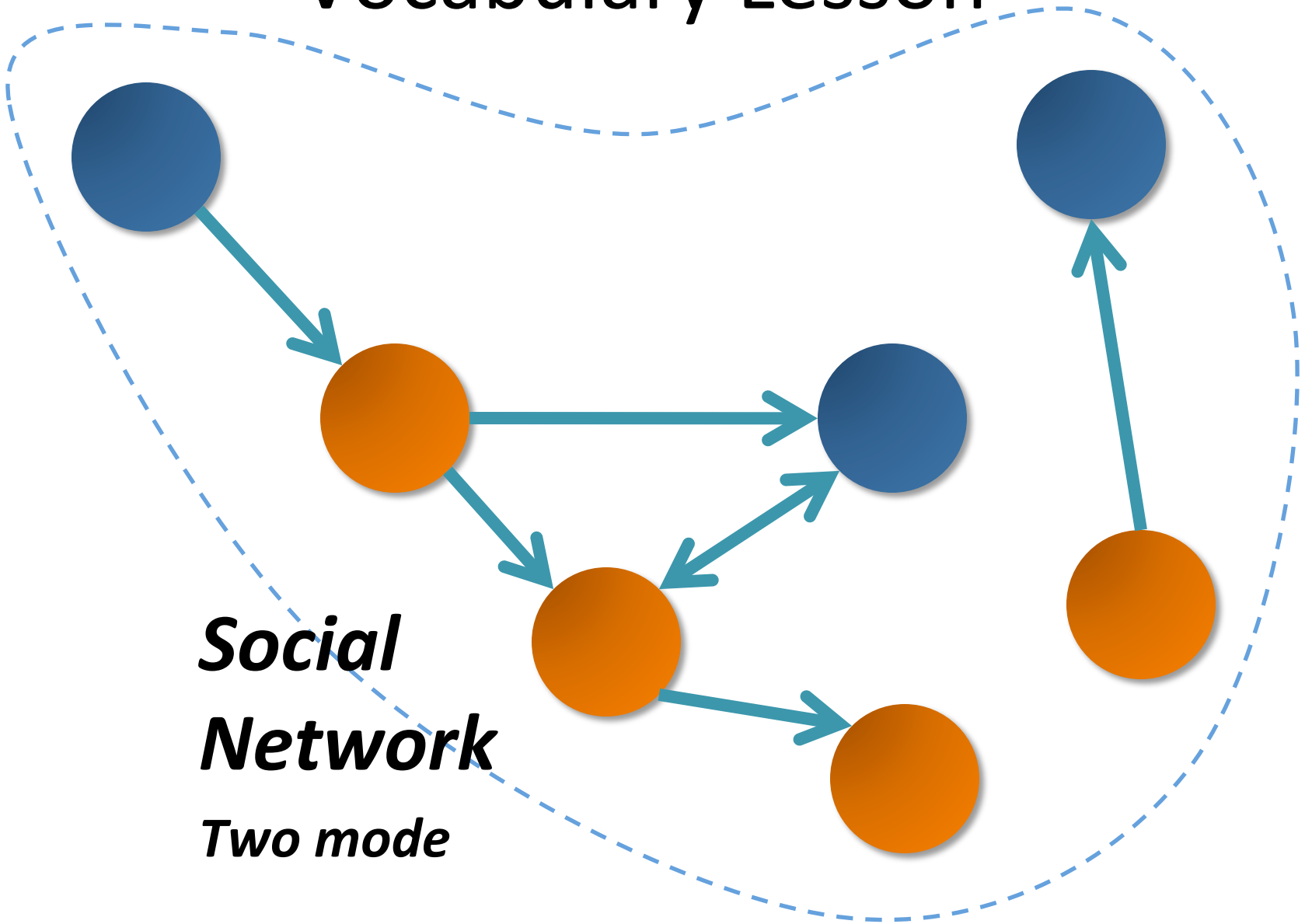
If A likes B and B likes C then A likes C (transitivity)
If A likes B and C likes B then A likes C
…



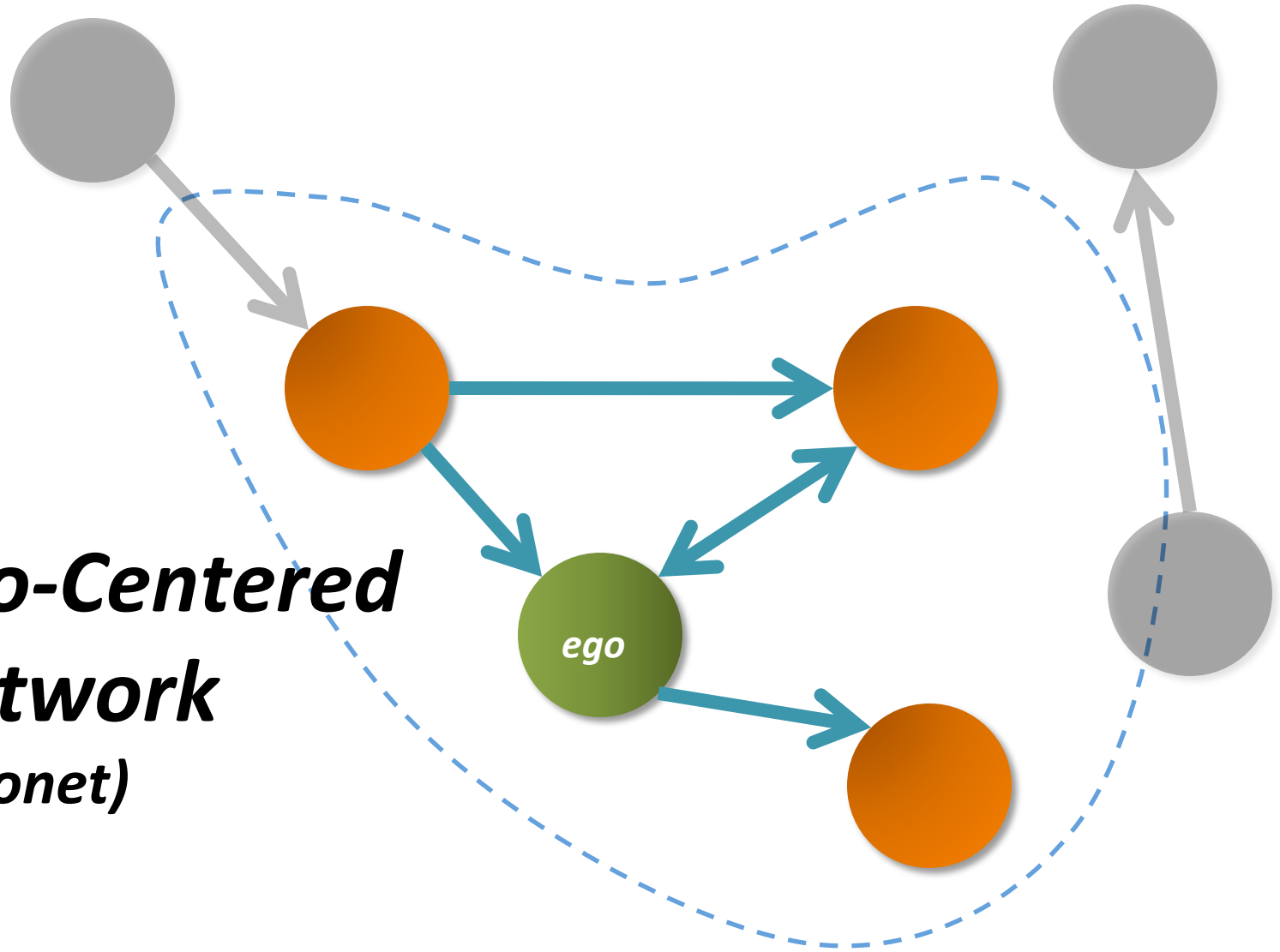
Vocabulary Lesson



Vocabulary Lesson



Vocabulary Lesson



Ego-Centered Network

(egonet)

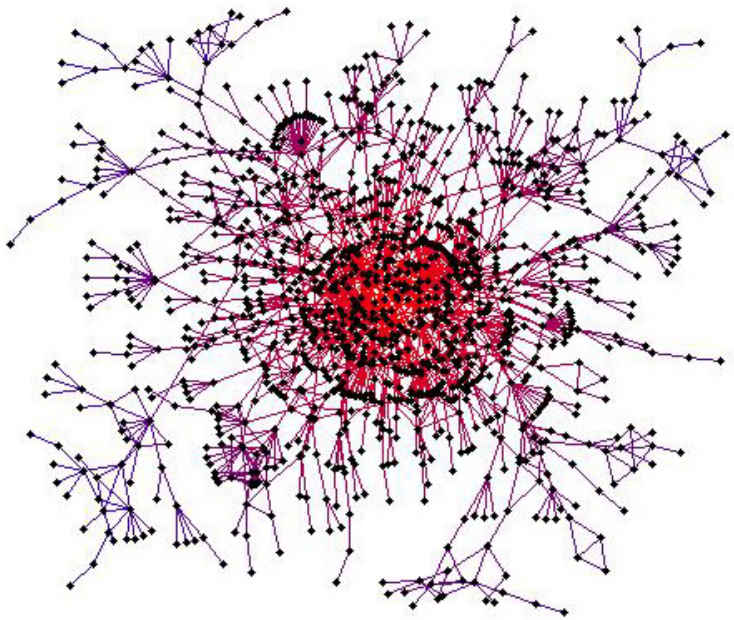
Describing Networks

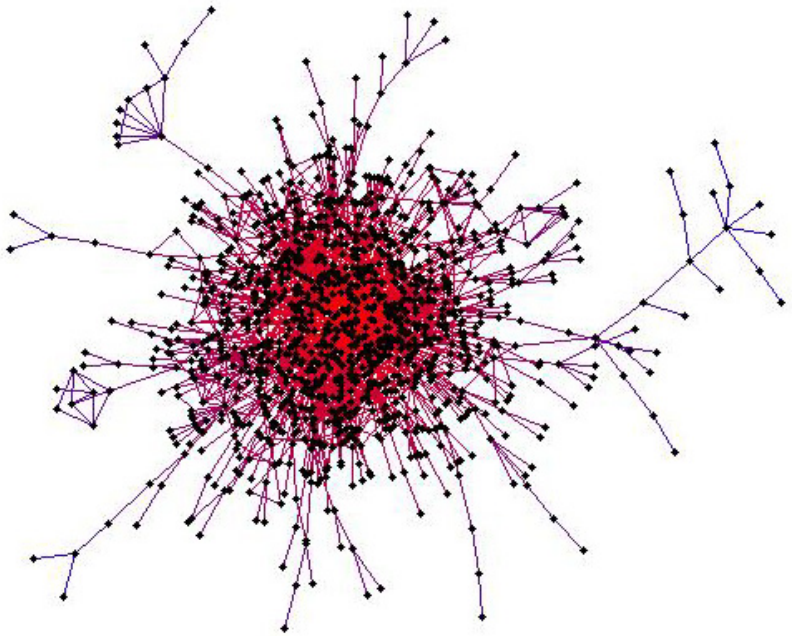
- Graph theoretic
  - Nodes/edges
- Sociometric
  - Sociomatrix (2D matrix representation)
  - Sociogram (the adjacency matrix)
- Algebraic
  - $n_i \rightarrow n_j$
- Basically complimentary

Describing Networks

MIT

Stanford





Describing Networks

- *Geodesic*
  - shortest_path(n,m)
- *Diameter*
  - max(geodesic(n,m)) n,m actors in graph
- *Density*
  - Number of existing edges / All possible edges
- *Degree distribution*

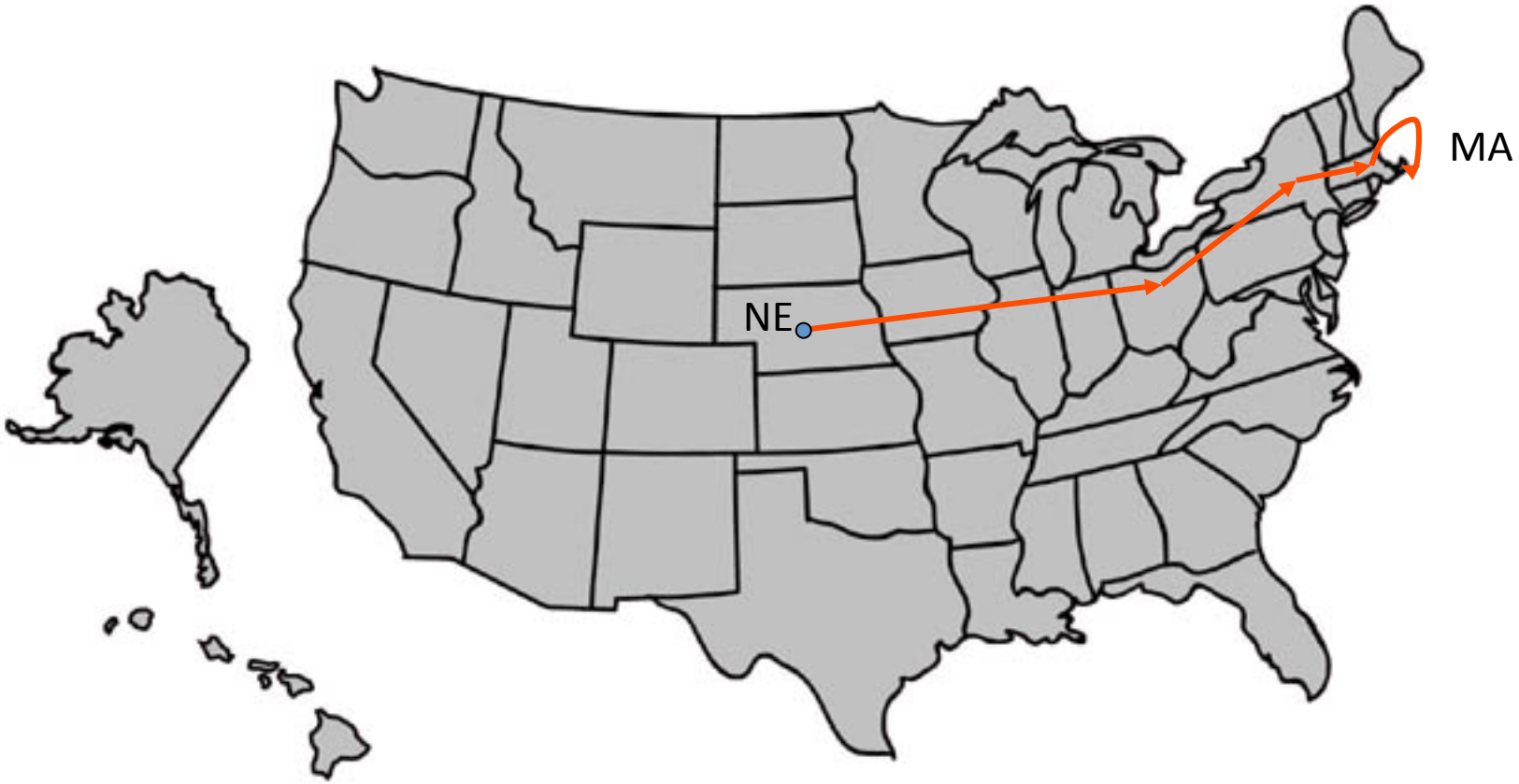
Types of Networks/Models

- A few quick examples
  - Erdős–Rényi
    - *G(n,M)*: randomly draw *M* edges between *n* nodes
    - Does not really model the real world
      - Average connectivity on nodes conserved

Types of Networks/Models

- A few quick examples
  - Erdős–Rényi
  - Small World
    - Watts-Strogatz
    - Kleinberg lattice model

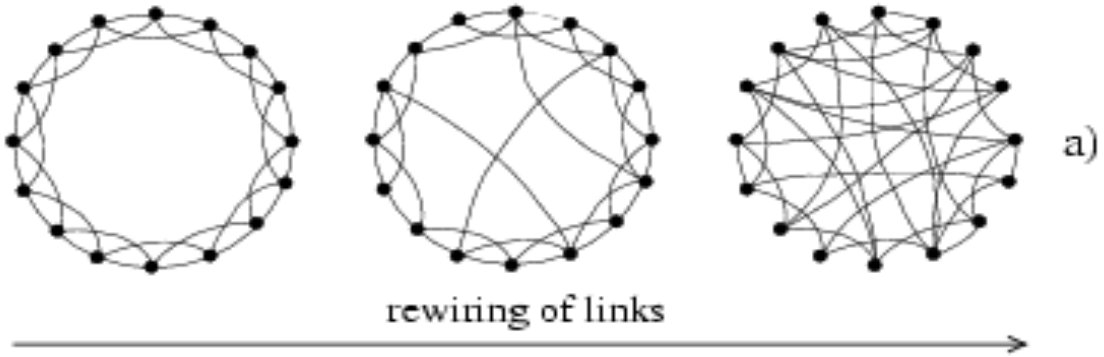
Small world experiments then



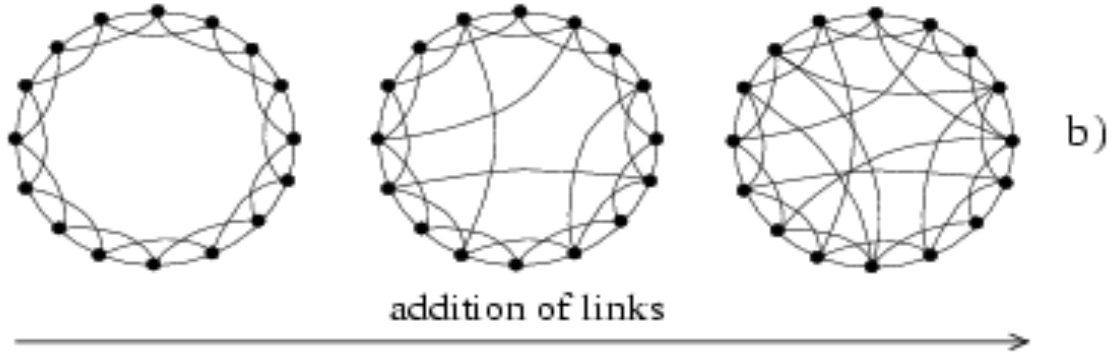
Milgram's experiment (1960's):

Given a target individual and a particular property, pass the message to a person you correspond with who is "closest" to the target.

Watts-Strogatz Ring Lattice Rewiring



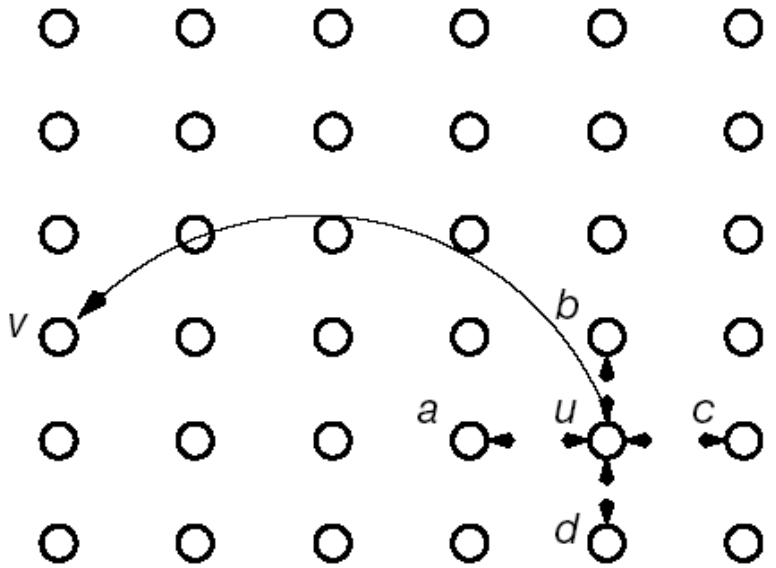
Select a fraction p of edges
Reposition on of their endpoints



Add a fraction p of additional edges leaving underlying lattice intact

- As in many network generating algorithms
  - Disallow self-edges
  - Disallow multiple edges

Kleinberg Lattice Model



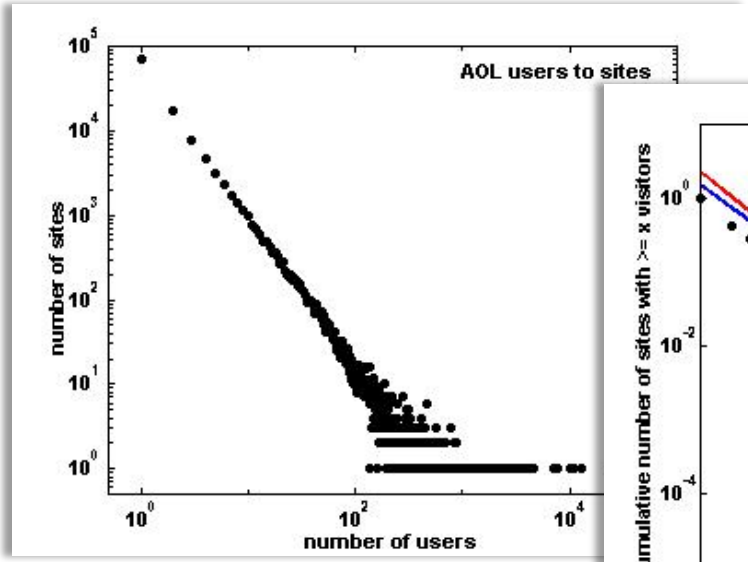
nodes are placed on a lattice and connect to nearest neighbors

additional links placed with $p_{uv} \sim d_{uv}^{-r}$

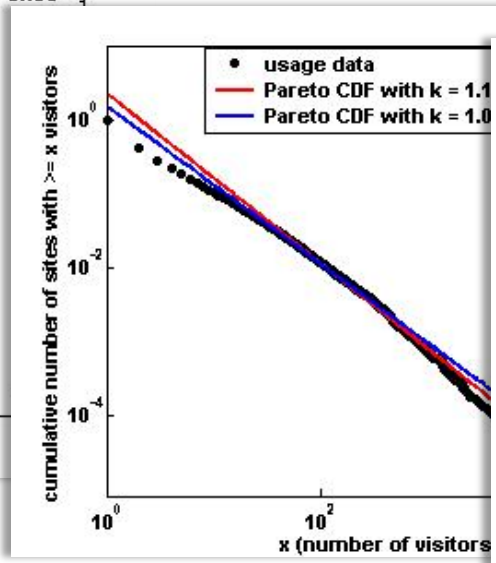
Kleinberg, 'The Small World Phenomenon, An Algorithmic Perspective' (Nature 2000)

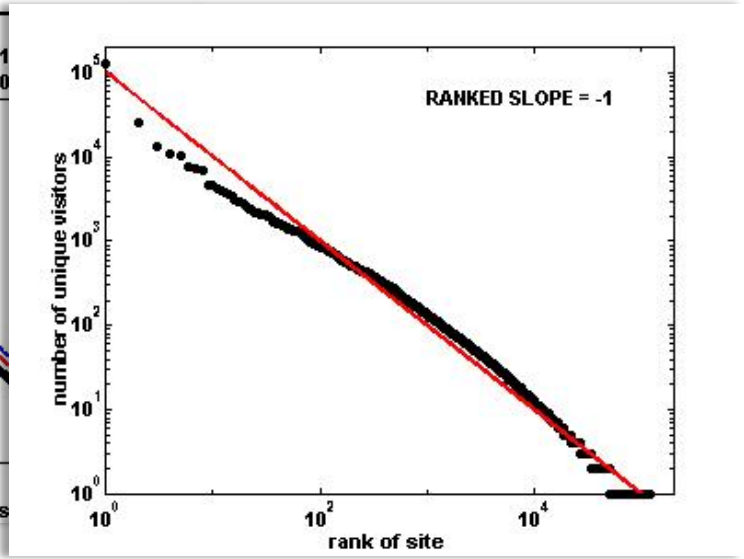
A little more on degree distribution

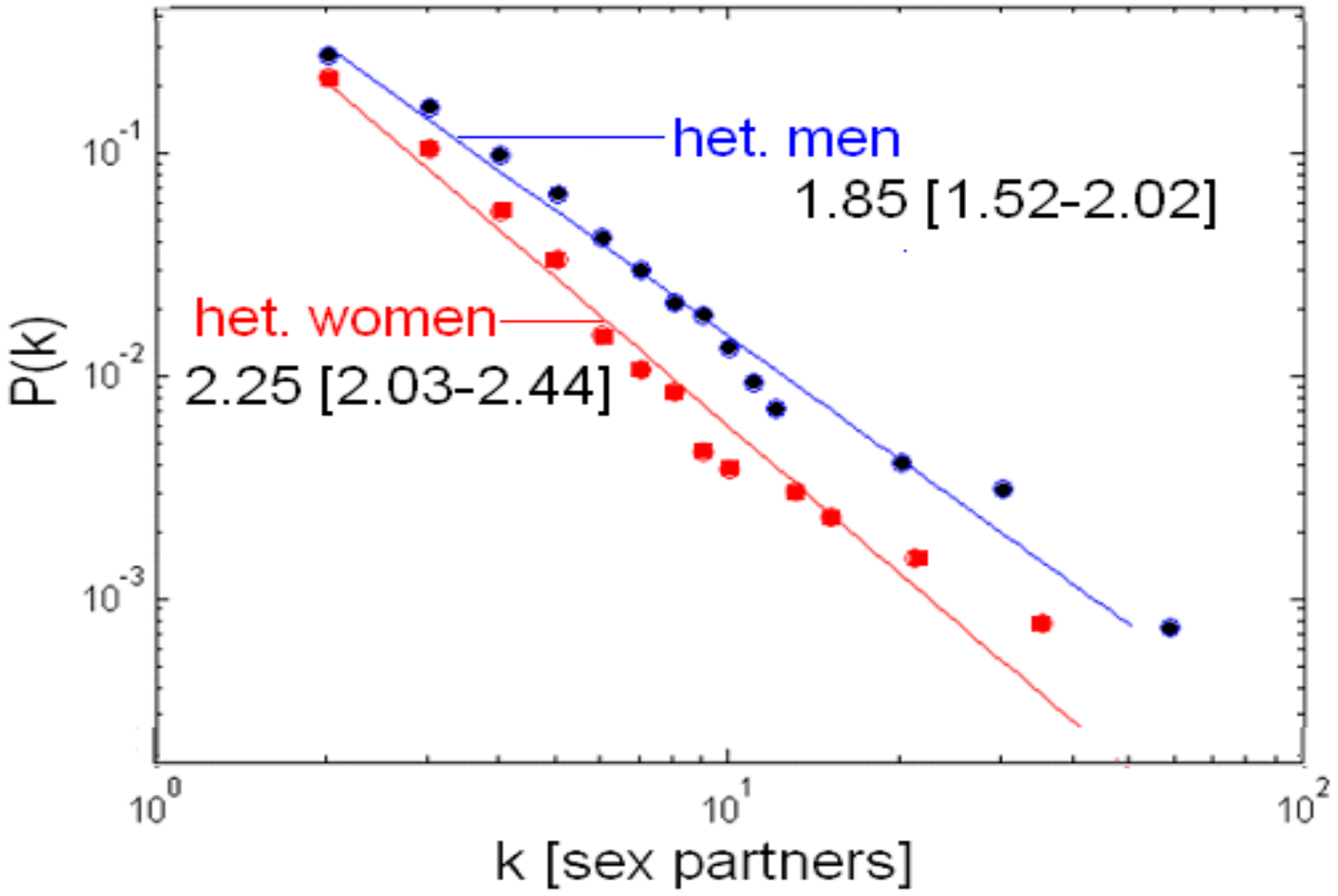
- Power-laws, zipf, etc.



CDF of users to sites






het. men
1.85 [1.52-2.02]
het. women
2.25 [2.03-2.44]
P(k)
10^{-1}
10^{-2}
10^{-3}
10^{0}
10^{1}
10^{2}
k [sex partners]

A little more on degree distribution

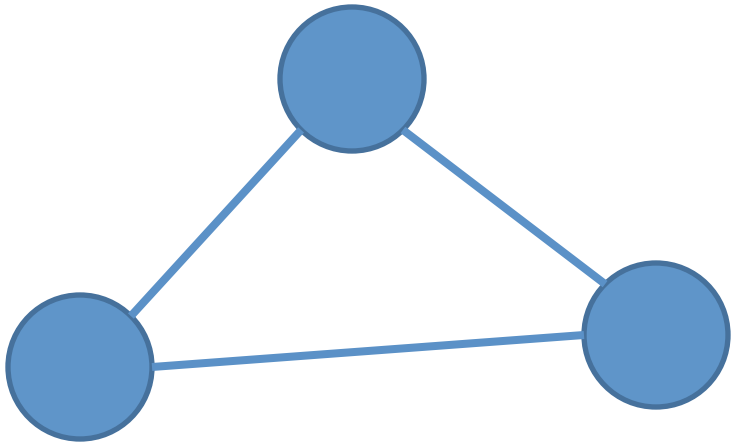
- Pareto/Power-law
  - Pareto: CDF $P[X > x] \sim x^{-k}$
  - Power-law: PDF $P[X = x] \sim x^{-(k+1)} = x^{-a}$
  - Some recent debate (Aaron Clauset)
    - http://arxiv.org/abs/0706.1062
- Zipf
  - Frequency versus rank $y \sim r^{-b}$ (small b)
- More info:
  - Zipf, Power-laws, and Pareto – a ranking tutorial (http://www.hpl.hp.com/research/idl/papers/ranking/ranking.html)

Types of Networks/Models

- A few quick examples
  - Erdős–Rényi
  - Small World
    - Watts-Strogatz
    - Kleinberg lattice model
  - Preferential Attachment
    - Generally attributed to Barabási & Albert

Basic BA-model

- Very simple algorithm to implement
  - start with an initial set of m_0 fully connected nodes
    - e.g. $m_0 = 3$



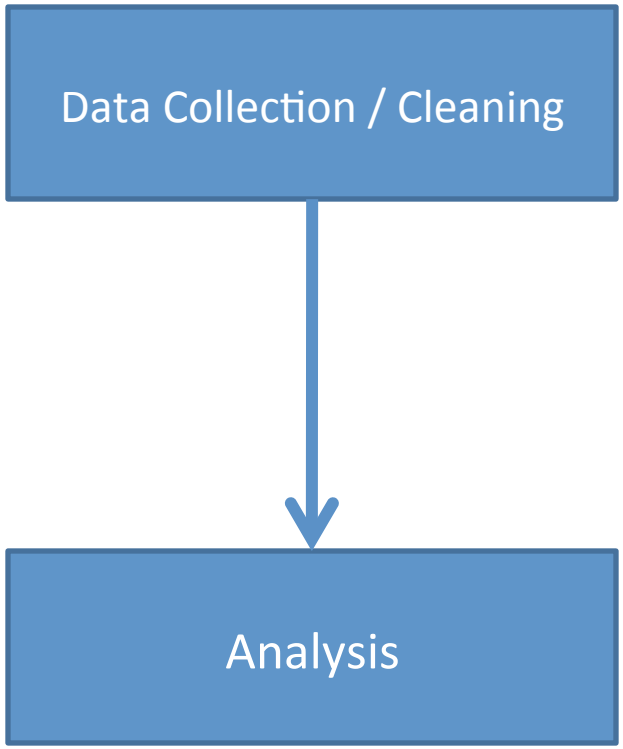
  - now add new vertices one by one, each one with exactly m edges
  - each new edge connects to an existing vertex in proportion to the number of edges that vertex already has → ***preferential attachment***

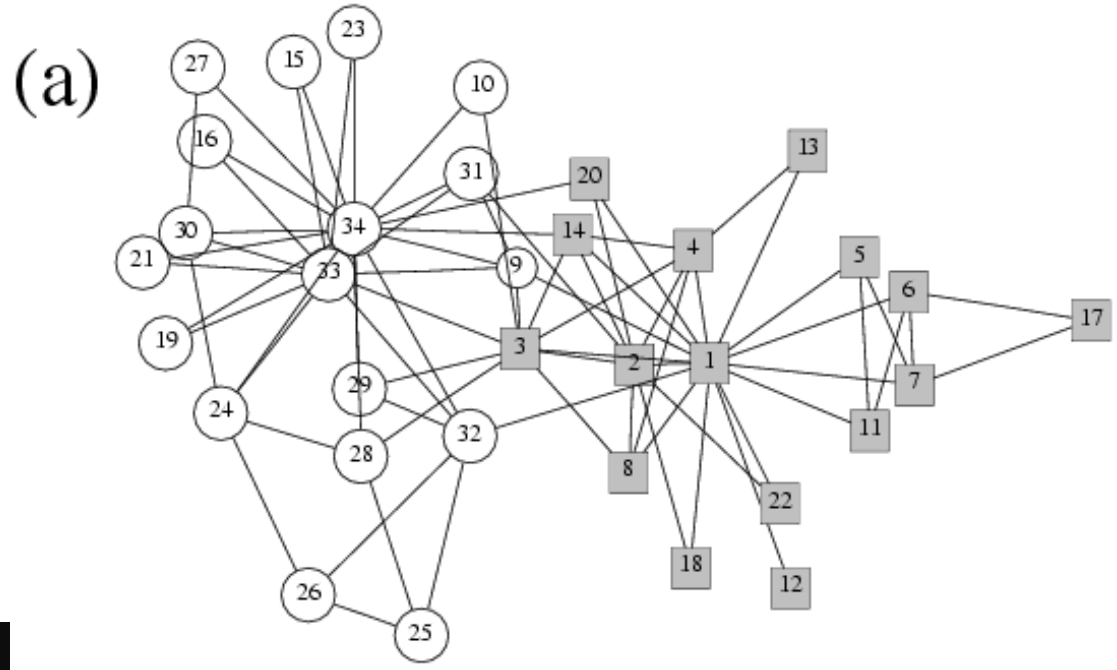
Properties of the BA graph

- The distribution is scale free with exponent $\alpha = 3$
  $P(k) = 2\ m^2/k^3$
- The graph is connected
  - Every new vertex is born with a link or several links (depending on whether m = 1 or m > 1)
  - It then connects to an 'older' vertex, which itself connected to another vertex when it was introduced
  - And we started from a connected core
- The older are richer
  - Nodes accumulate links as time goes on, which gives older nodes an advantage since newer nodes are going to attach preferentially – and older nodes have a higher degree to tempt them with than some new kid on the block

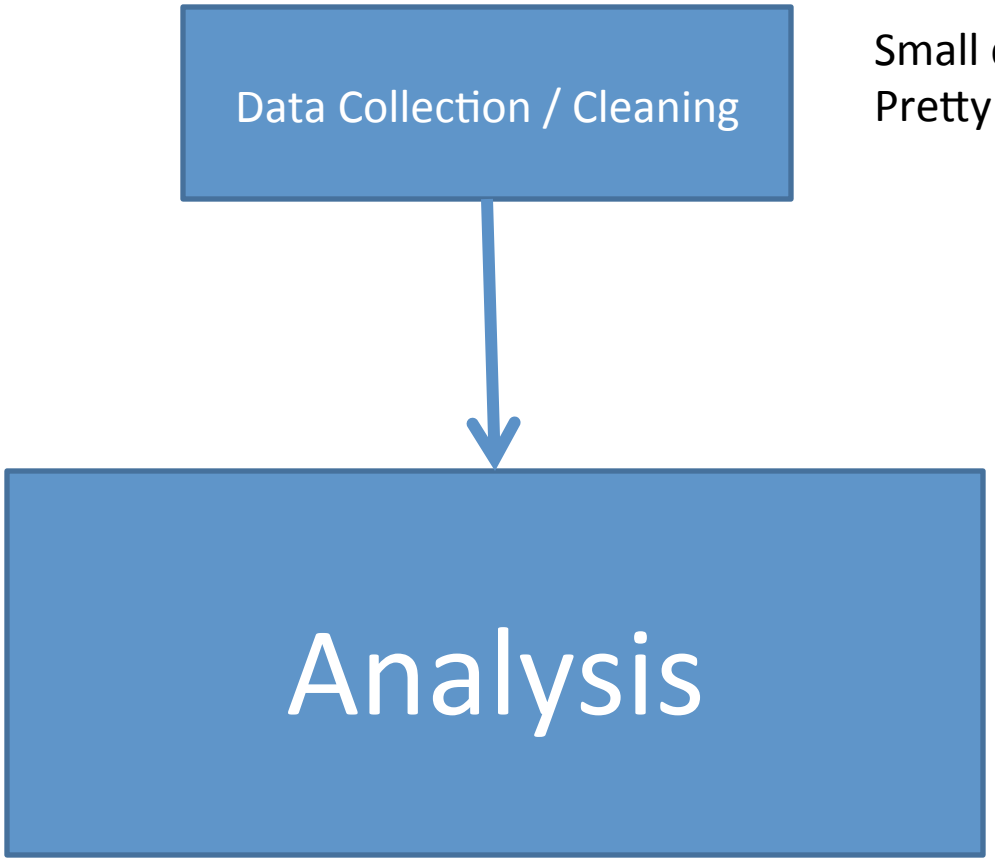
Common Tasks

- Measuring “importance”
  - Centrality, prestige
- Link prediction
- Diffusion modeling
  - Epidemiological
- Clustering
  - Blockmodeling, Girvan-Newman
- Structure analysis
  - Motifs, Isomorphisms, etc.
- Visualization/Privacy/etc.


Data Collection / Cleaning
Analysis


(a)
23
27
15
10
16
31
20
13
30
34
14
4
5
21
33
9
6
17
19
3
2
1
7
29
24
32
11
28
8
22
18
12
26
25

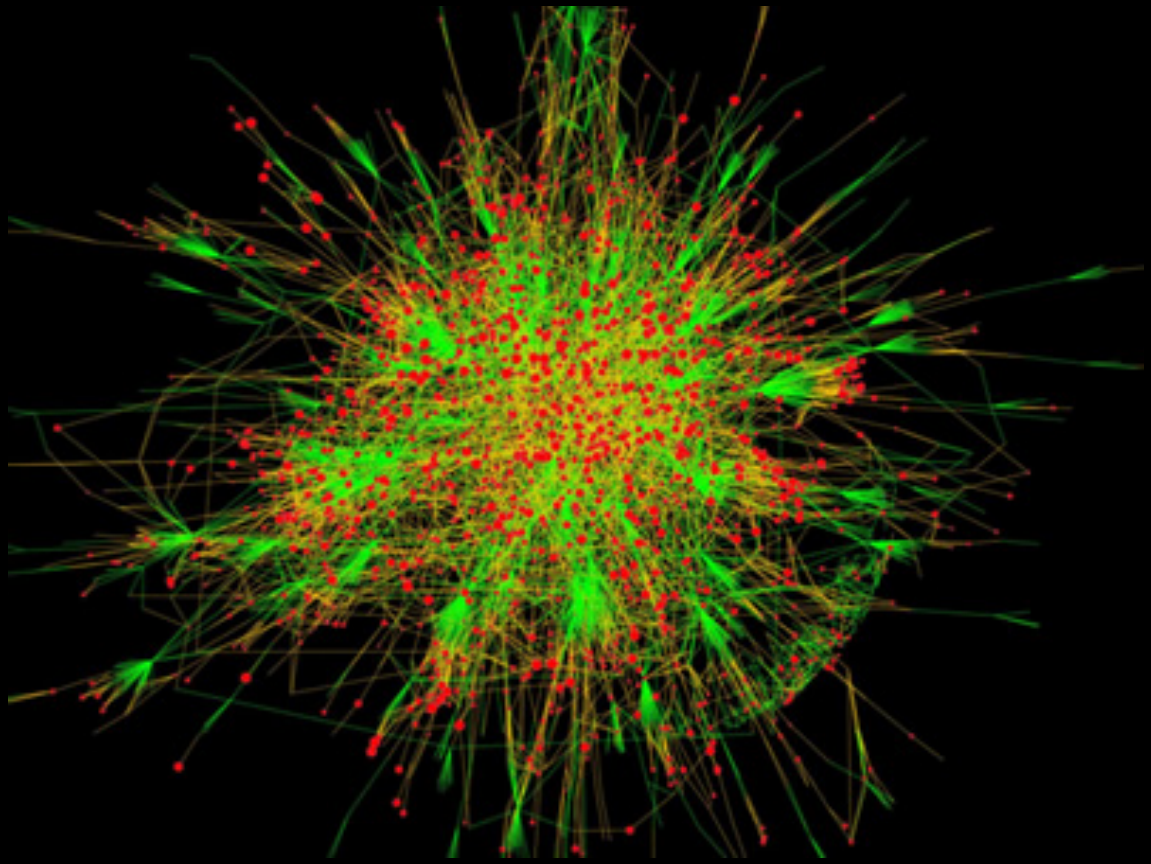
Past

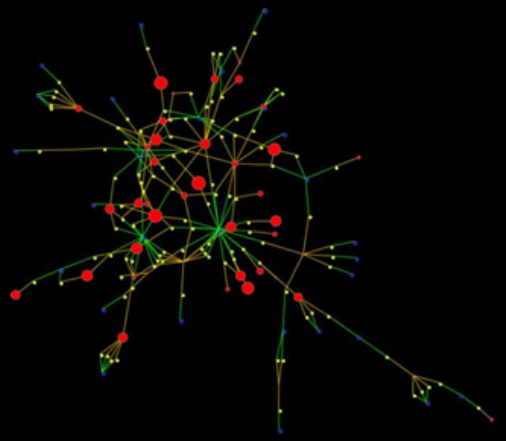


Small datasets
Pretty explicit connections

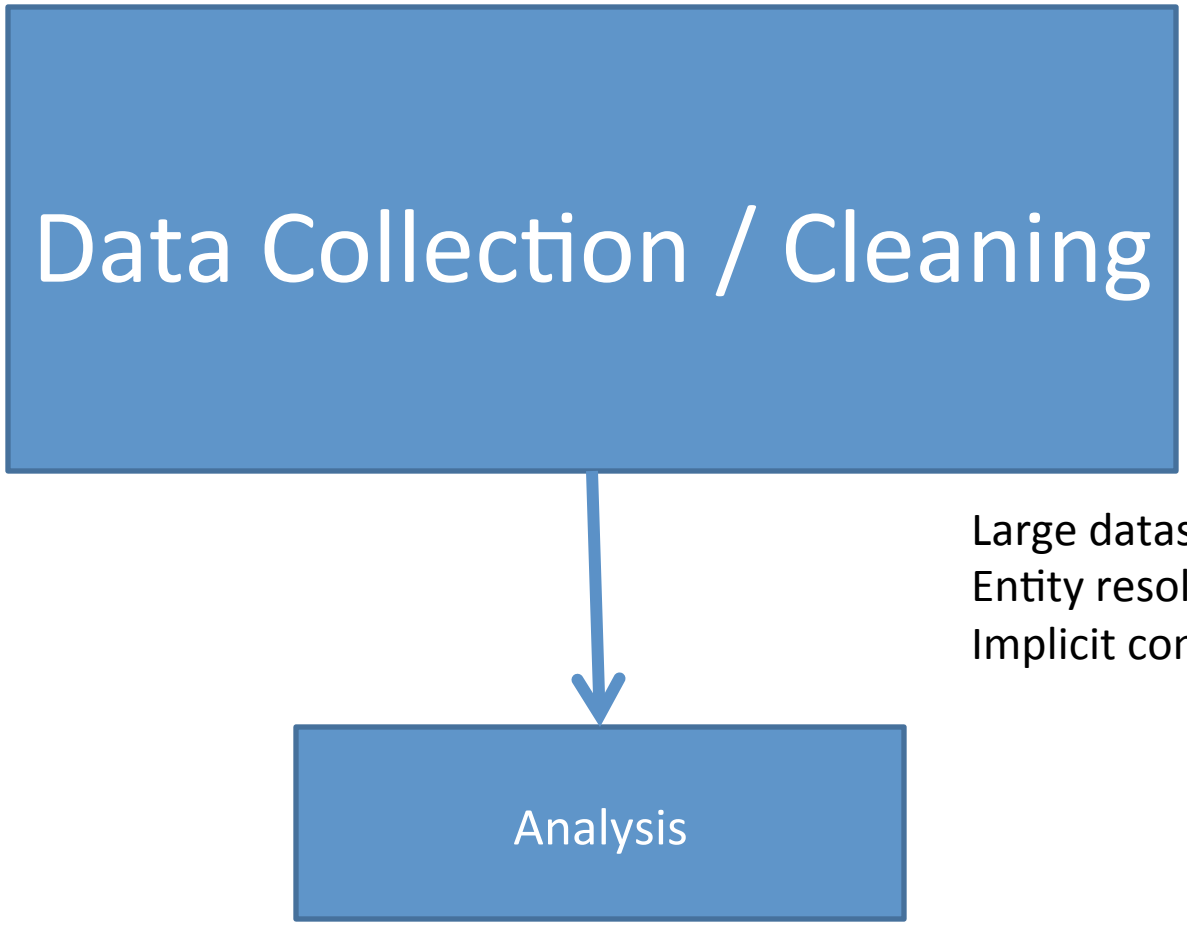
Understand the properties

Past





Present



Large datasets
Entity resolution
Implicit connections

Understand the properties

Present

Common Tasks

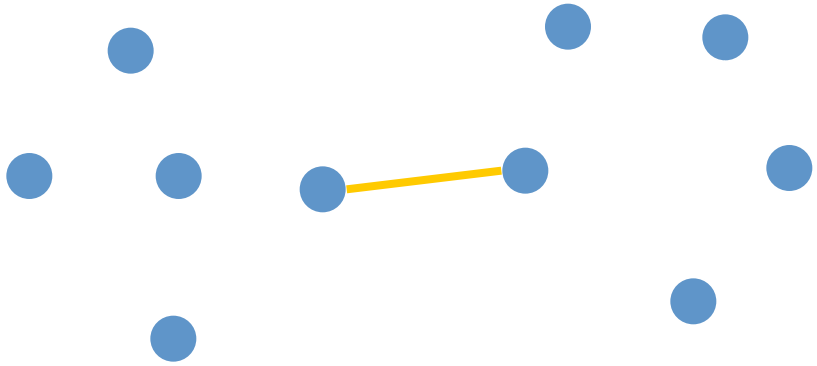
- Measuring "importance"
  - Centrality, prestige (incoming links)
- Link prediction
- Diffusion modeling
  - Epidemiological
- Clustering
  - Blockmodeling, Girvan-Newman
- Structure analysis
  - Motifs, Isomorphisms, etc.
- Visualization/Privacy/etc.

Centrality Measures

- Degree centrality
  - Edges per node (the more, the more important the node)
- Closeness centrality
  - How close the node is to every other node
- Betweenness centrality
  - How many shortest paths go through the edge node (communication metaphor)
- Information centrality
  - All paths to other nodes weighted by path length
- Bibliometric + Internet style
  - PageRank

Tie Strength

- Strength of Weak Ties (Granovetter)
  - Granovetter: How often did you see the contact that helped you find the job prior to the job search
    - 16.7 % often (at least once a week)
    - 55.6% occasionally (more than once a year but less than twice a week)
    - 27.8% rarely – once a year or less
  - Weak ties will tend to have different *information* than we and our close contacts do

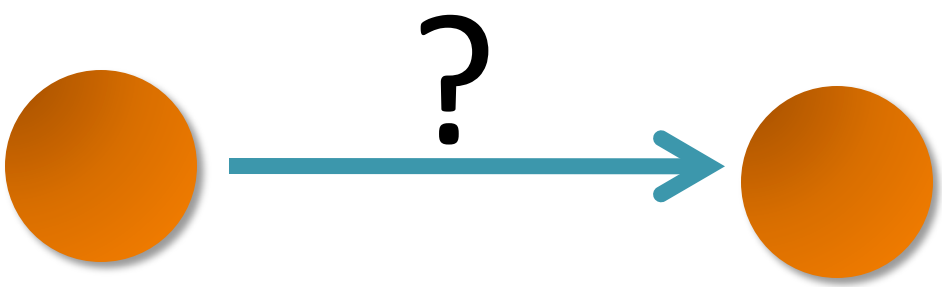


weak ties will tend to have high beweenness and low transitivity

Common Tasks

- Measuring "importance"
  - Centrality, prestige (incoming links)
- Link prediction
- Diffusion modeling
  - Epidemiological
- Clustering
  - Blockmodeling, Girvan-Newman
- Structure analysis
  - Motifs, Isomorphisms, etc.
- Visualization/Privacy/etc.

Link Prediction



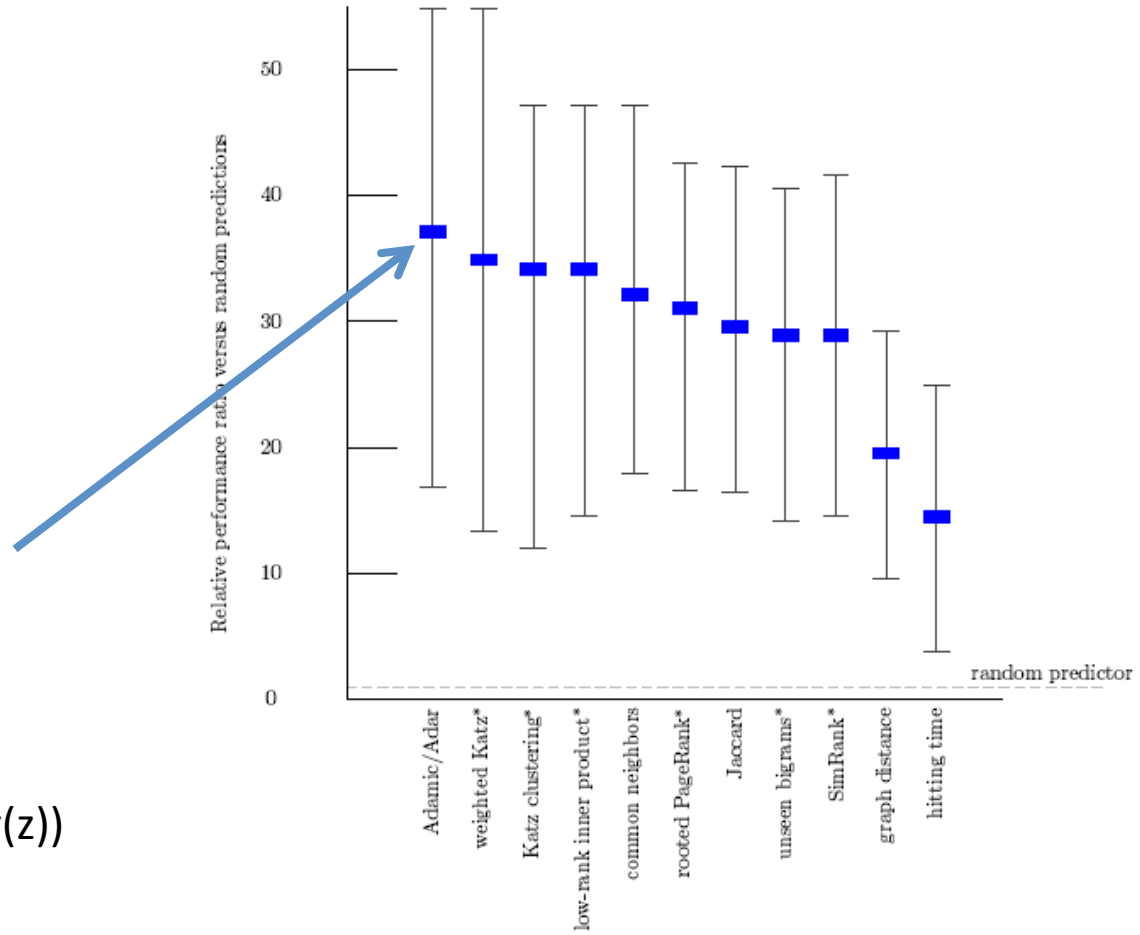
Link Prediction in Social Net Data

- *We know things about structure*
  - *Homophily* = like likes like or bird of a feather flock together or similar people group together
  - *Mutuality*
  - *Triad closure*

- Various measures that try to use this

Link Prediction

- Simple metrics
  - Only take into account graph properties

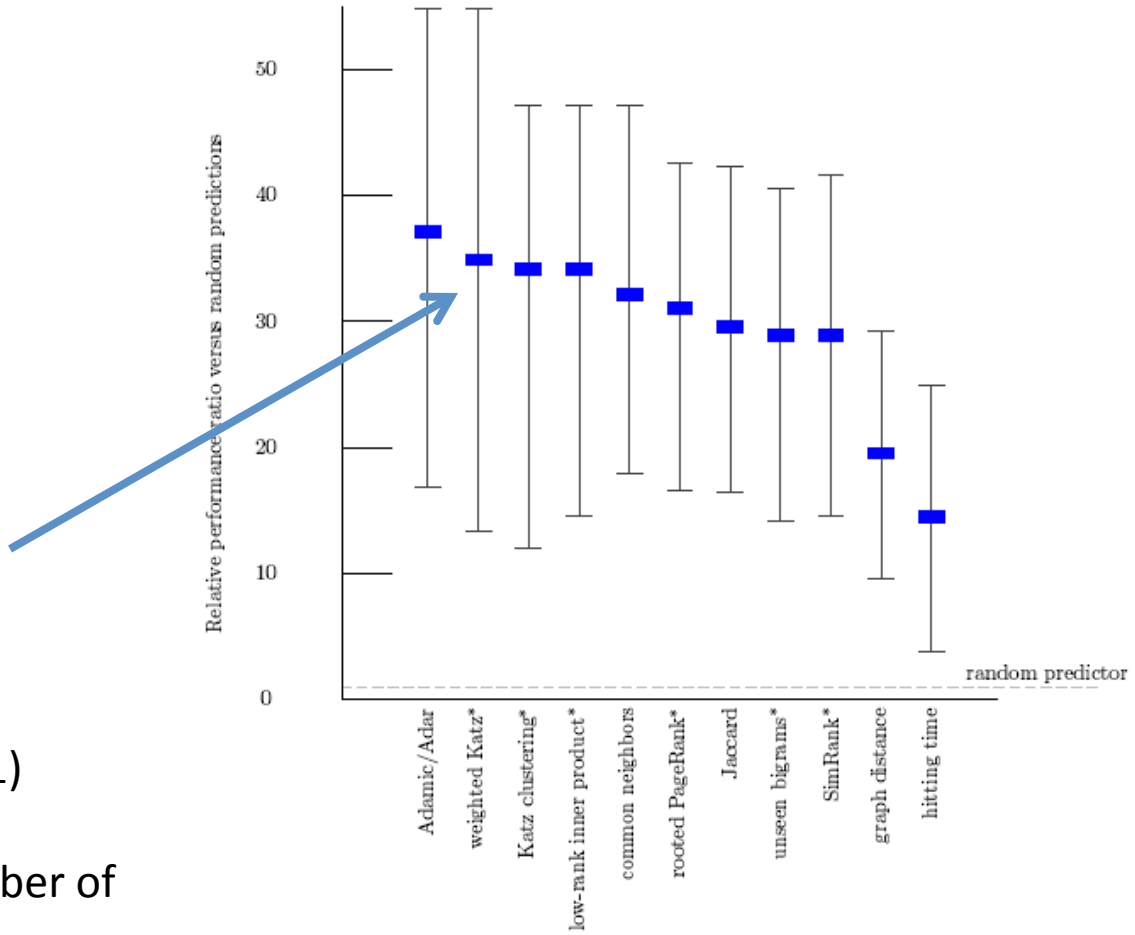


$\Gamma(x)$ = neighbors of x
Originally: 1 / log(frequency(z))

Liben-Nowell, Kleinberg (CIKM'03)

Link Prediction

- Simple metrics
  - Only take into account graph properties



Paths of length *l* (generally 1)
from x to y
weighted variant is the number of times the two collaborated

Liben-Nowell, Kleinberg (CIKM'03)

Link Prediction in Relational Data

- *We know things about structure*
  - *Homophily* = like likes like or bird of a feather flock together or similar people group together
  - *Mutuality*
  - *Triad closure*

- Slightly more interesting problem if we have relational data on actors and ties
  - Move beyond structure

Relationship & Link Prediction

Employee /
contractor
Salary
Time at company

Common Tasks

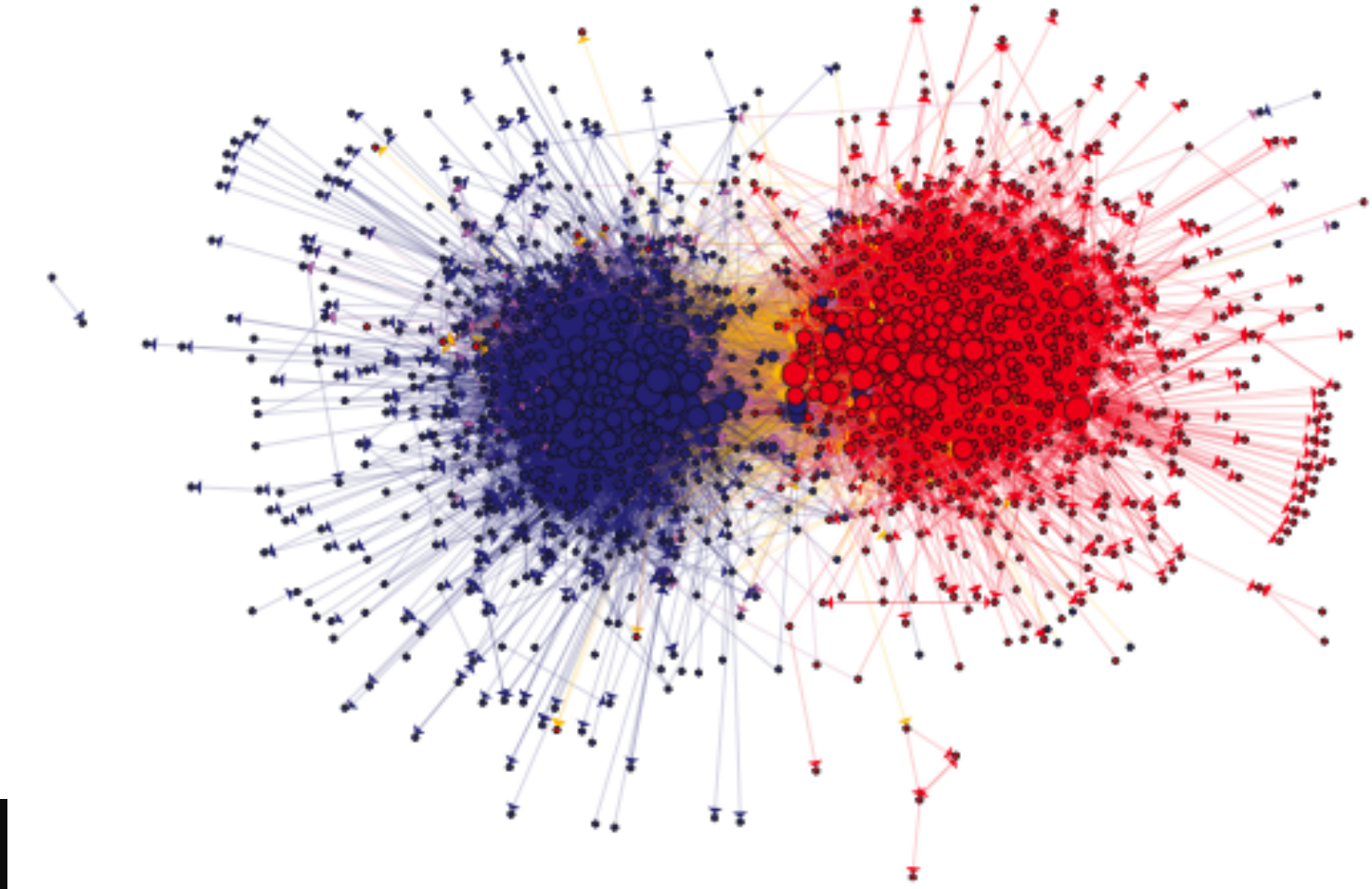
- Measuring "importance"
  - Centrality, prestige (incoming links)
- Link prediction
- Diffusion modeling
  - Epidemiological
- Clustering
  - Blockmodeling, Girvan-Newman
- Structure analysis
  - Motifs, Isomorphisms, etc.
- Visualization/Privacy/etc.

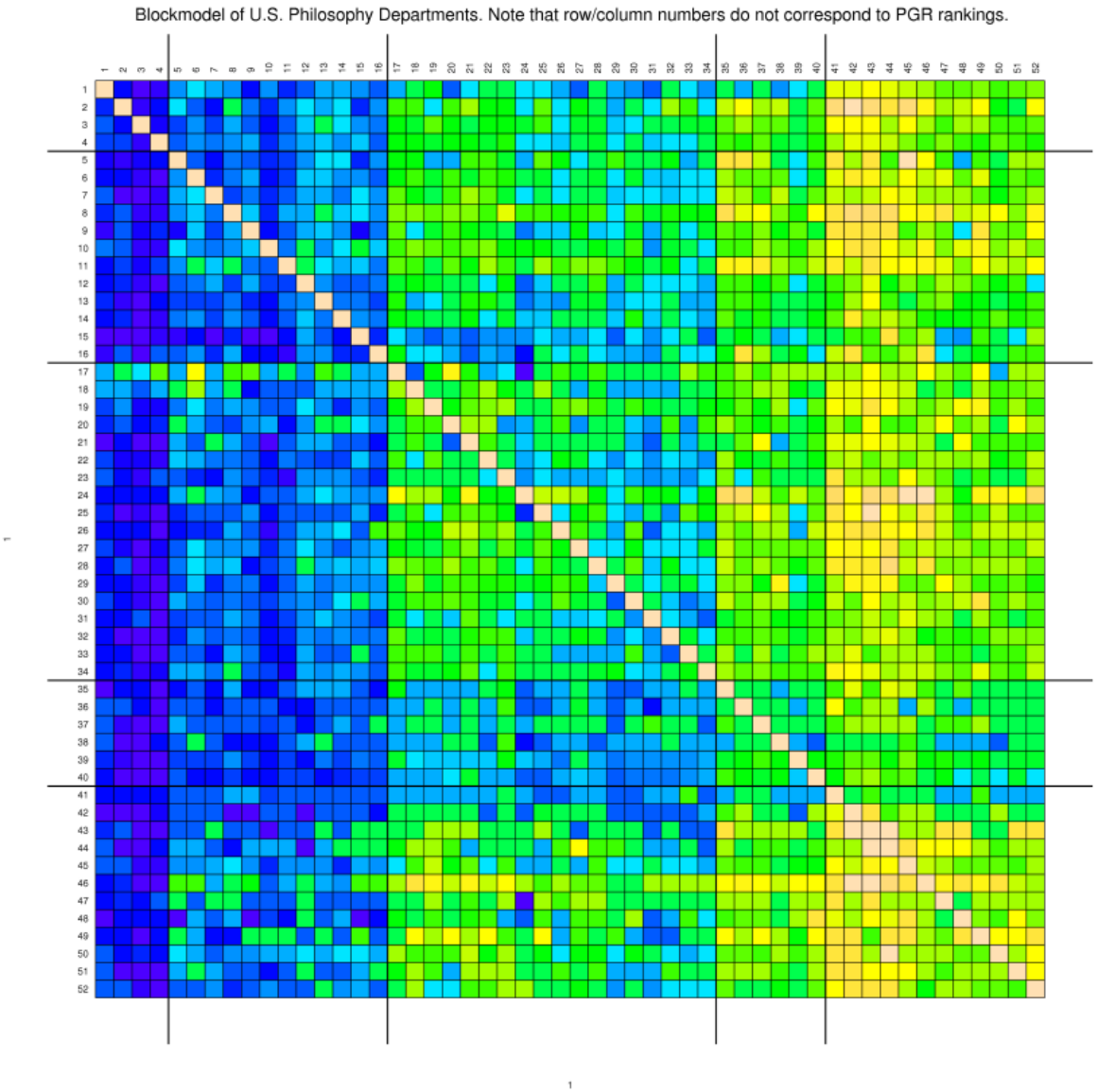
Epidemiological

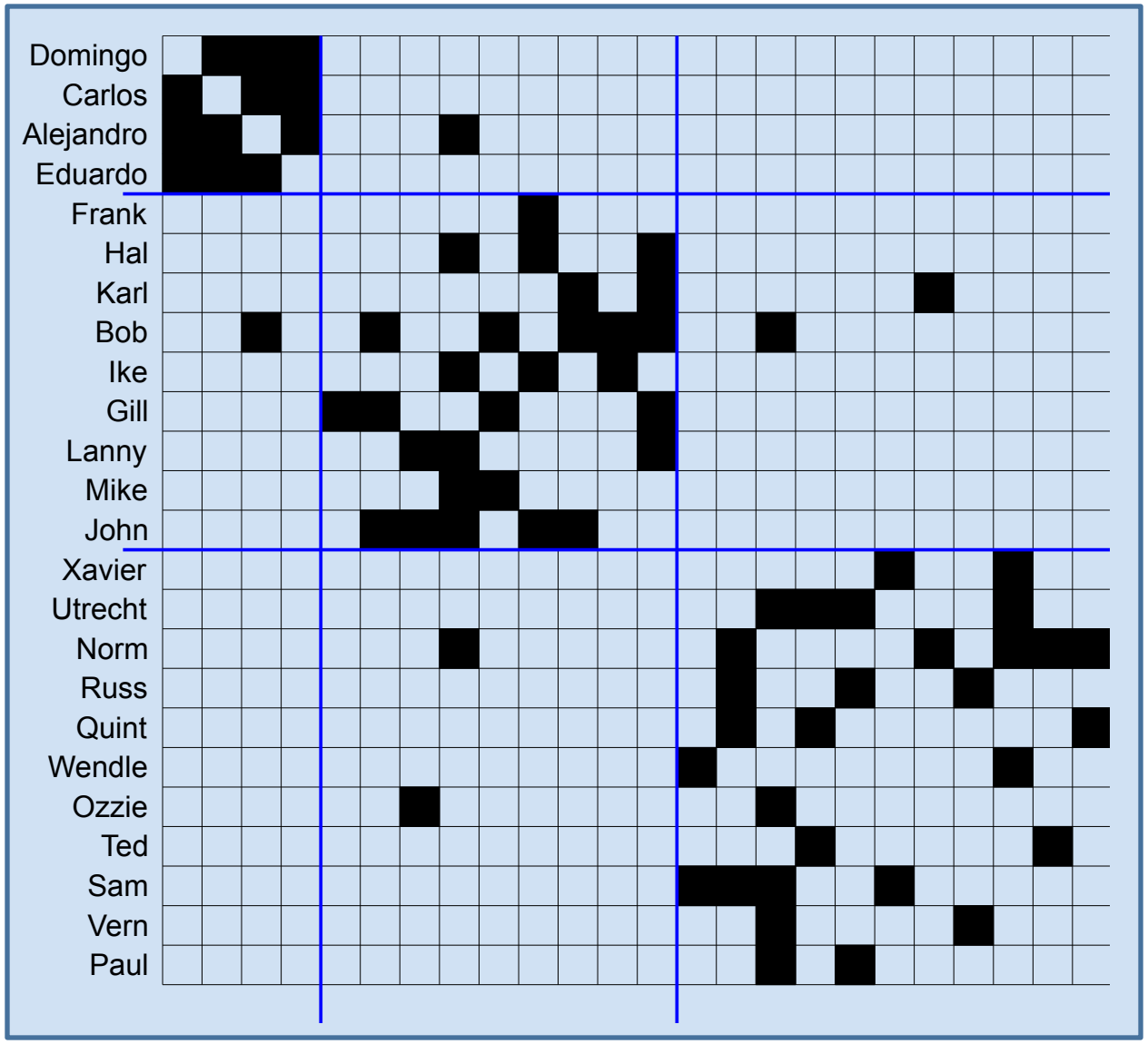
- Viruses
  - Biological, computational
  - STDs, needle sharing, etc.
  - Mark Handcock at UW
- Blog networks
  - Applying SIR models (Info Diffusion Through Blogspace, Gruhl et al.)
    - Induce transmission graph, cascade models, simulation
  - Link prediction (Tracking Information Epidemics in Blogspace, Adar et al.)
    - Find repeated “likely” infections
  - Outbreak detection (Cost-effective Outbreak Detection in Networks, Leskovec et al.)
    - Submodularity

Common Tasks

- Measuring "importance"
  - Centrality, prestige (incoming links)
- Link prediction
- Diffusion modeling
  - Epidemiological
- Clustering/Community Detection
  - Blockmodeling, Girvan-Newman
- Structure analysis
  - Motifs, Isomorphisms, etc.
- Visualization/Privacy/etc.




Blockmodel of U.S. Philosophy Departments. Note that row/column numbers do not correspond to PGR rankings.
1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52

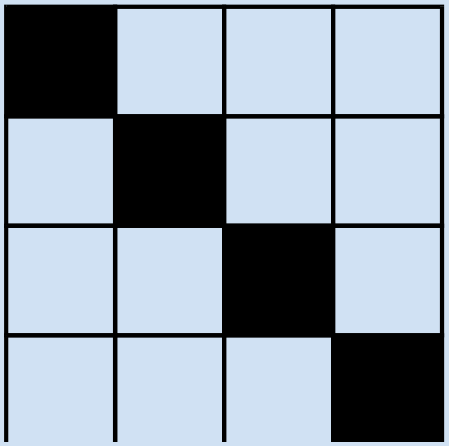

Domingo
Carlos
Alejandro
Eduardo
Frank
Hal
Karl
Bob
Ike
Gill
Lanny
Mike
John
Xavier
Utrecht
Norm
Russ
Quint
Wendle
Ozzie
Ted
Sam
Vern
Paul

Blockmodels

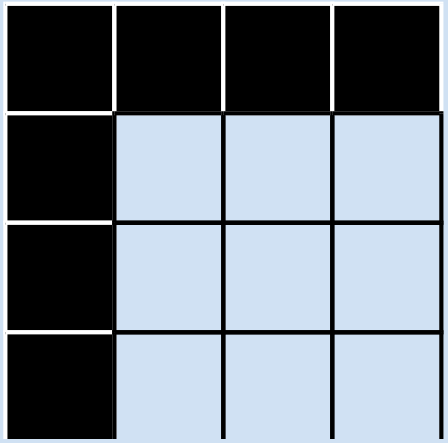
- Actors are portioned into *positions*
  - Rearrange rows/columns
- The sociomatrix is then reduced to a smaller *image*
- Hierarchical clustering
  - Various distance metrics
    - Euclidean, CONvergence of CORrelation (CONCOR)
- Various "fit" metrics

Image matrix

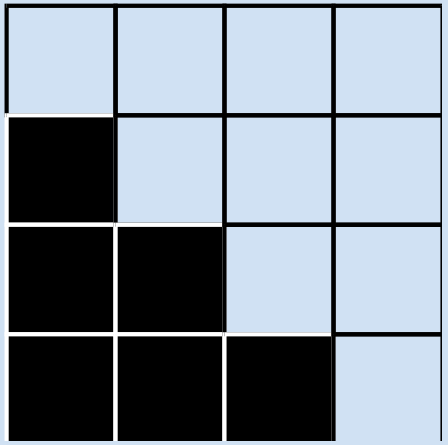
Cohesion

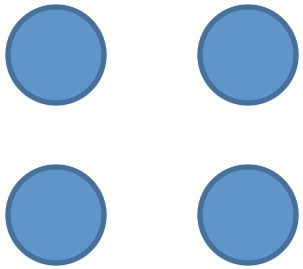


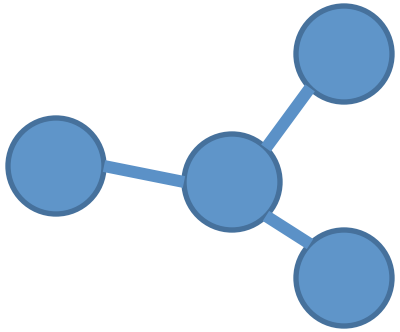
Center-periphery

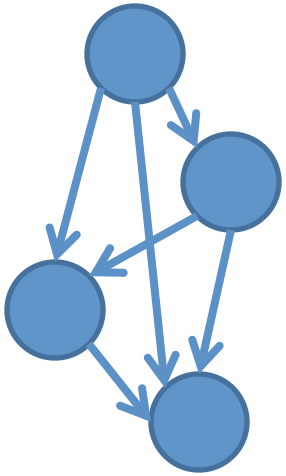


Ranking









Girvan-Newman Algorithm

- Split on shortest paths ("weak ties")

1. Calculate betweenness on all edges
2. Remove highest betweenness edge
3. Recalculate
4. Goto 1

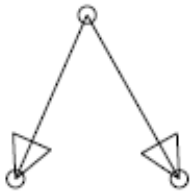
Other solutions

- Min-cut based
- “Voltage” based
- Hierarchical schemes

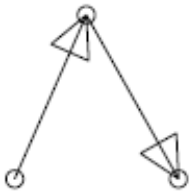
Common Tasks

- Measuring "importance"
  - Centrality, prestige (incoming links)
- Link prediction
- Diffusion modeling
  - Epidemiological
- Clustering
  - Blockmodeling, Girvan-Newman
- Structure analysis
  - Motifs, Isomorphisms, etc.
- Visualization/Privacy/etc.

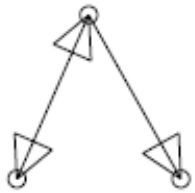
id6



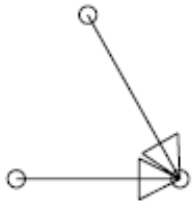
id12



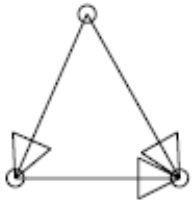
id14



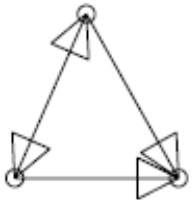
id36



id38

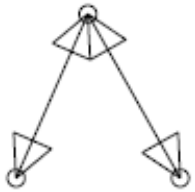


id46

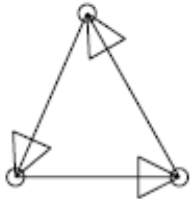


id74

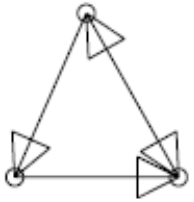
id78



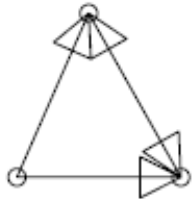
id98



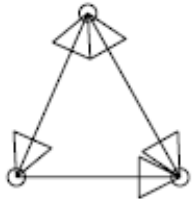
id102



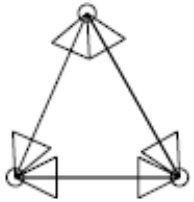
id108

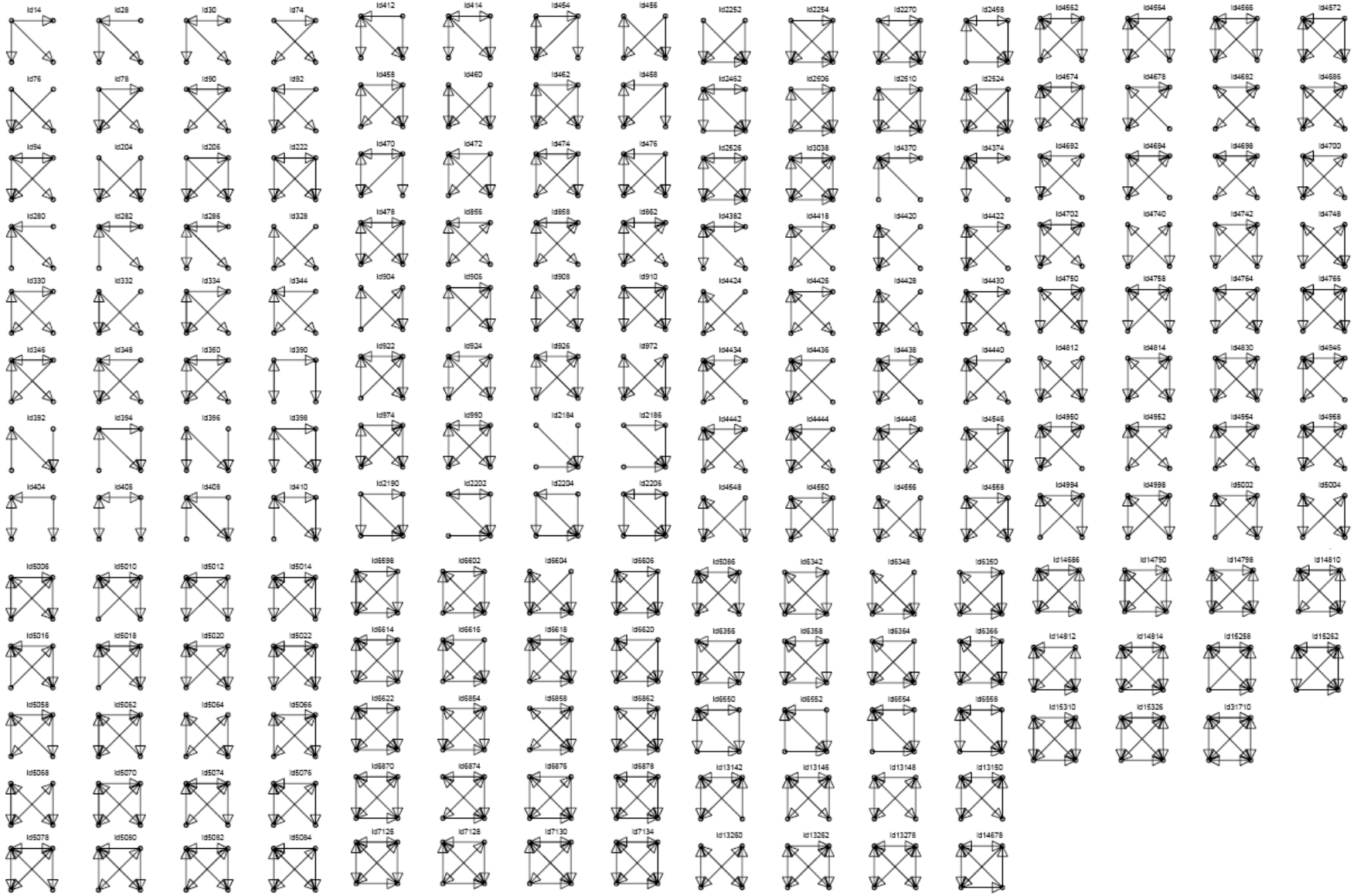


id110



id238




Id14
Id28
Id30
Id74
Id412
Id414
Id454
Id456
Id2252
Id2254
Id2270
Id2468
Id4552
Id4554
Id4566
Id4572
Id76
Id78
Id90
Id92
Id458
Id460
Id462
Id468
Id2462
Id2506
Id2510
Id2524
Id4574
Id4678
Id4682
Id4686
Id94
Id204
Id206
Id222
Id470
Id472
Id474
Id476
Id2526
Id3038
Id4370
Id4374
Id4692
Id4694
Id4698
Id4700
Id280
Id282
Id286
Id328
Id478
Id856
Id858
Id862
Id4382
Id4418
Id4420
Id4422
Id4702
Id4740
Id4742
Id4748
Id330
Id332
Id334
Id344
Id904
Id905
Id908
Id910
Id4424
Id4426
Id4428
Id4430
Id4750
Id4758
Id4764
Id4766
Id346
Id348
Id350
Id390
Id922
Id924
Id926
Id972
Id4434
Id4436
Id4438
Id4440
Id4812
Id4814
Id4830
Id4946
Id392
Id394
Id396
Id398
Id974
Id990
Id2184
Id2186
Id4442
Id4444
Id4446
Id4546
Id4950
Id4952
Id4954
Id4968
Id404
Id406
Id408
Id410
Id2190
Id2202
Id2204
Id2206
Id4548
Id4550
Id4566
Id4568
Id4994
Id4998
Id5002
Id5004
Id5006
Id5010
Id5012
Id5014
Id6598
Id6602
Id6604
Id6606
Id5086
Id5342
Id5348
Id5350
Id14686
Id14790
Id14798
Id14810
Id5016
Id5018
Id5020
Id5022
Id6614
Id6616
Id6618
Id6620
Id6356
Id5358
Id5364
Id5366
Id14812
Id14814
Id15258
Id15262
Id5058
Id5062
Id5064
Id5066
Id6622
Id6854
Id6858
Id6862
Id6550
Id6552
Id6554
Id6558
Id15310
Id15326
Id31710
Id5068
Id5070
Id5074
Id5076
Id6870
Id6874
Id6876
Id6878
Id13142
Id13146
Id13148
Id13150
Id5078
Id5080
Id5082
Id5084
Id7126
Id7128
Id7130
Id7134
Id13260
Id13262
Id13278
Id14678

Network motif detection

- How many more motifs of a certain type exist over a random network

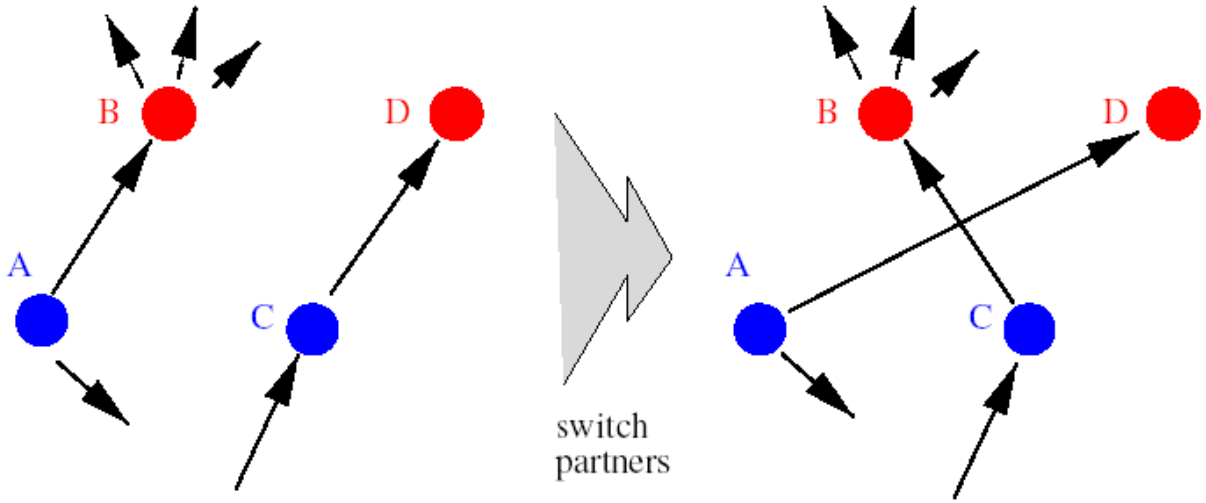
- Started in biological networks
  - http://www.weizmann.ac.il/mcb/UriAlon/

Basic idea

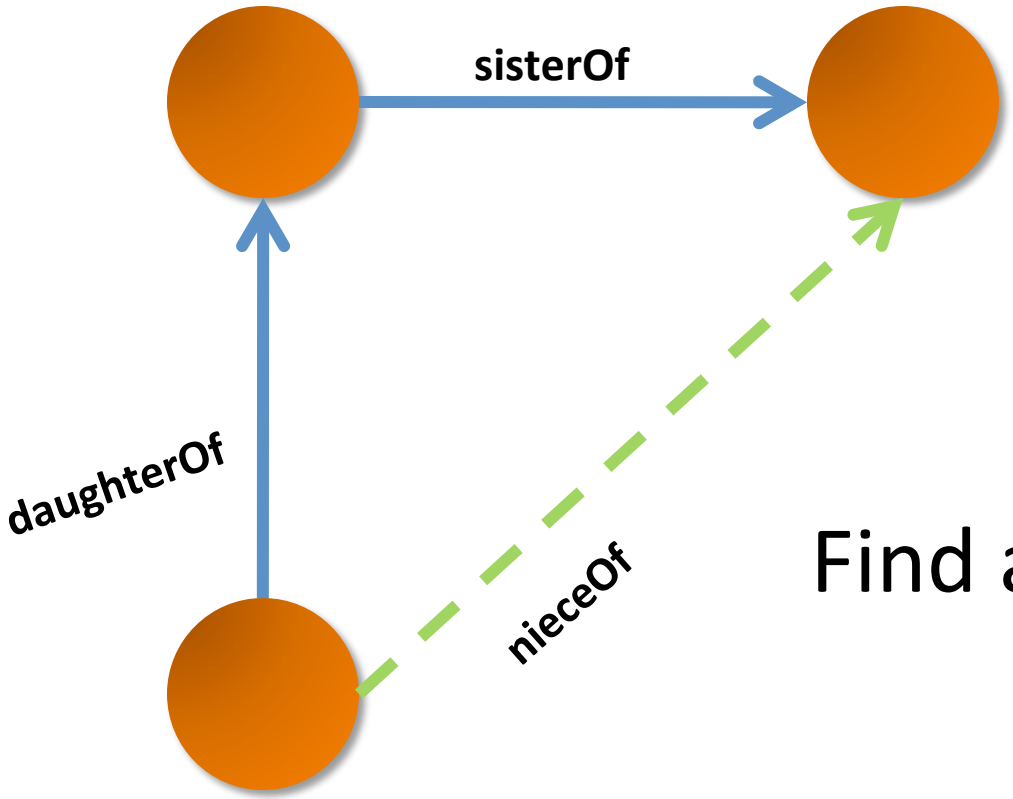
- construct many random graphs with the same number of nodes and edges (same node degree distribution?)
- count the number of motifs in those graphs
- calculate the Z score: the probability that the given number of motifs in the real world network could have occurred by chance

Generating random graphs

- Many models don't preserve the desired features
- Have to be careful how we generate



Other Structural Analysis



Find all

Common Tasks

- Measuring "importance"
  - Centrality, prestige (incoming links)
- Link prediction
- Diffusion modeling
  - Epidemiological
- Clustering
  - Blockmodeling, Girvan-Newman
- Structure analysis
  - Motifs, Isomorphisms, etc.
- Visualization/Privacy/etc.

Privacy

- Emerging interest in anonymizing networks
  - Lars Backstrom (WWW'07) demonstrated one of the first attacks
- How to remove labels while preserving graph properties?
  - While ensuring that labels cannot be reapplied

Software

- Pajek
  - http://vlado.fmf.uni-lj.si/pub/networks/pajek/
- UCINET
  - http://www.analytictech.com/
- KrackPlot
  - http://www.andrew.cmu.edu/user/krack/krackplot.shtml
- GUESS
  - http://www.graphexploration.org
- Etc.

Books/Journals/Conferences

- *Social Networks/Phs. Rev*
- Social Network Analysis (Wasserman + Faust)
- The Development of Social Network Analysis (Freeman)
- Linked (Barabsi)
- Six Degrees (Watts)
- Sunbelt/ICWSM/KDD/CIKM/NIPS

Assortativity

- Social networks are assortative:
  - the gregarious people associate with other gregarious people
  - the loners associate with other loners
- The Internet is disassorative: