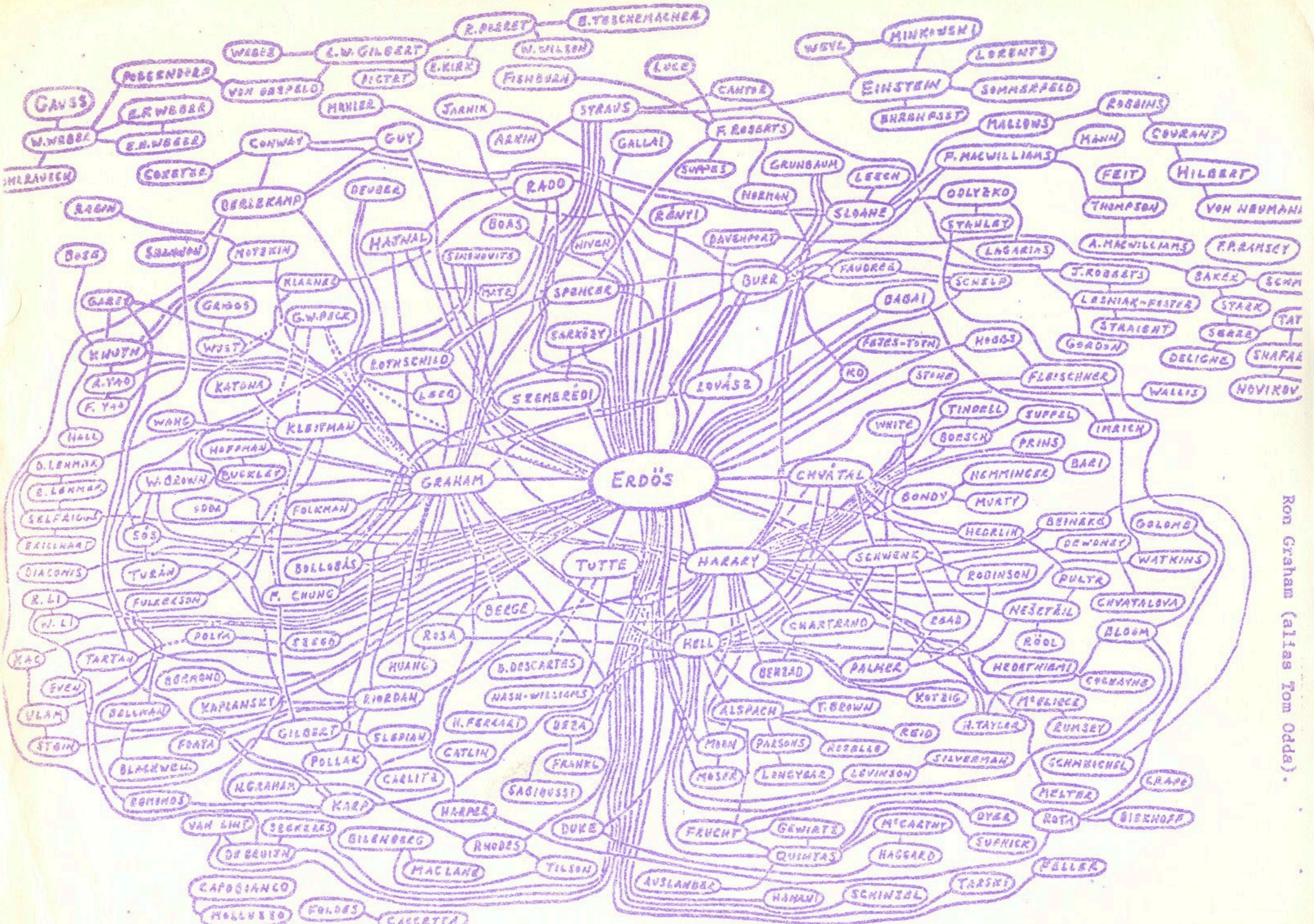
Social Network Analysis

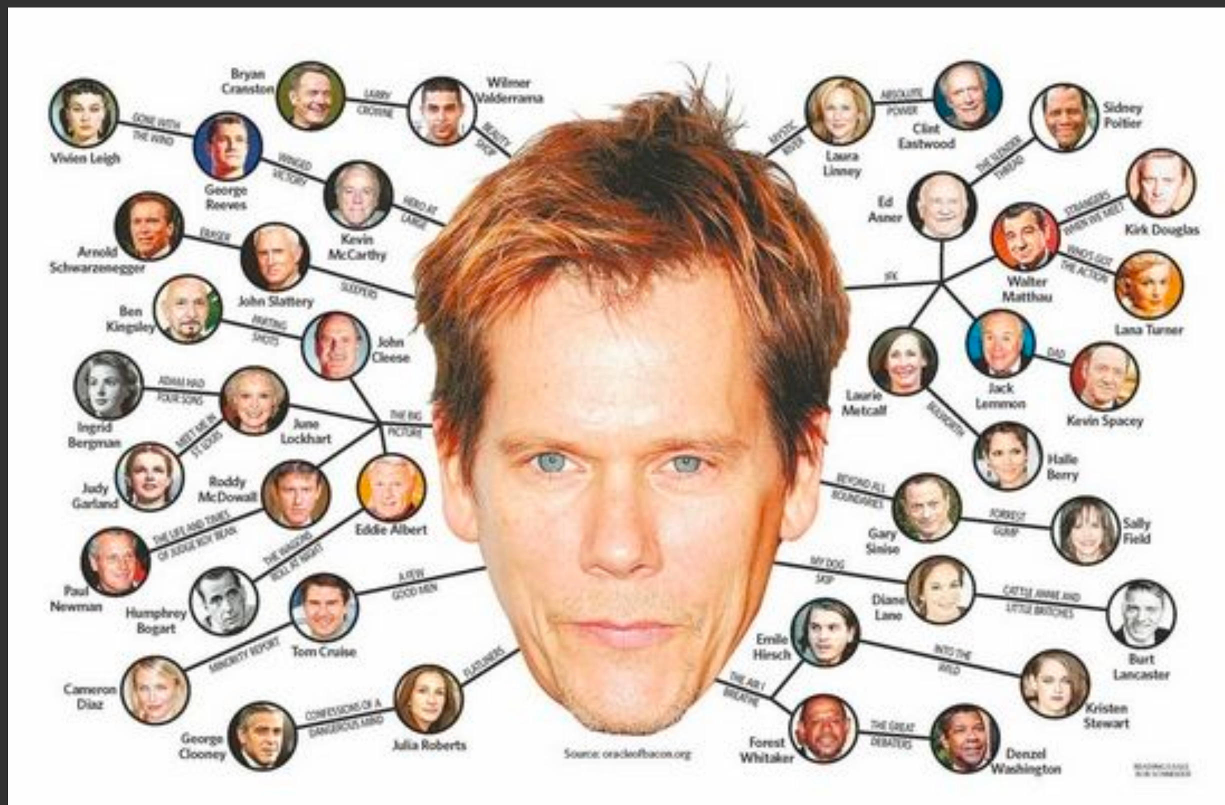
Some content from Lada Adamic and Eytan Adar



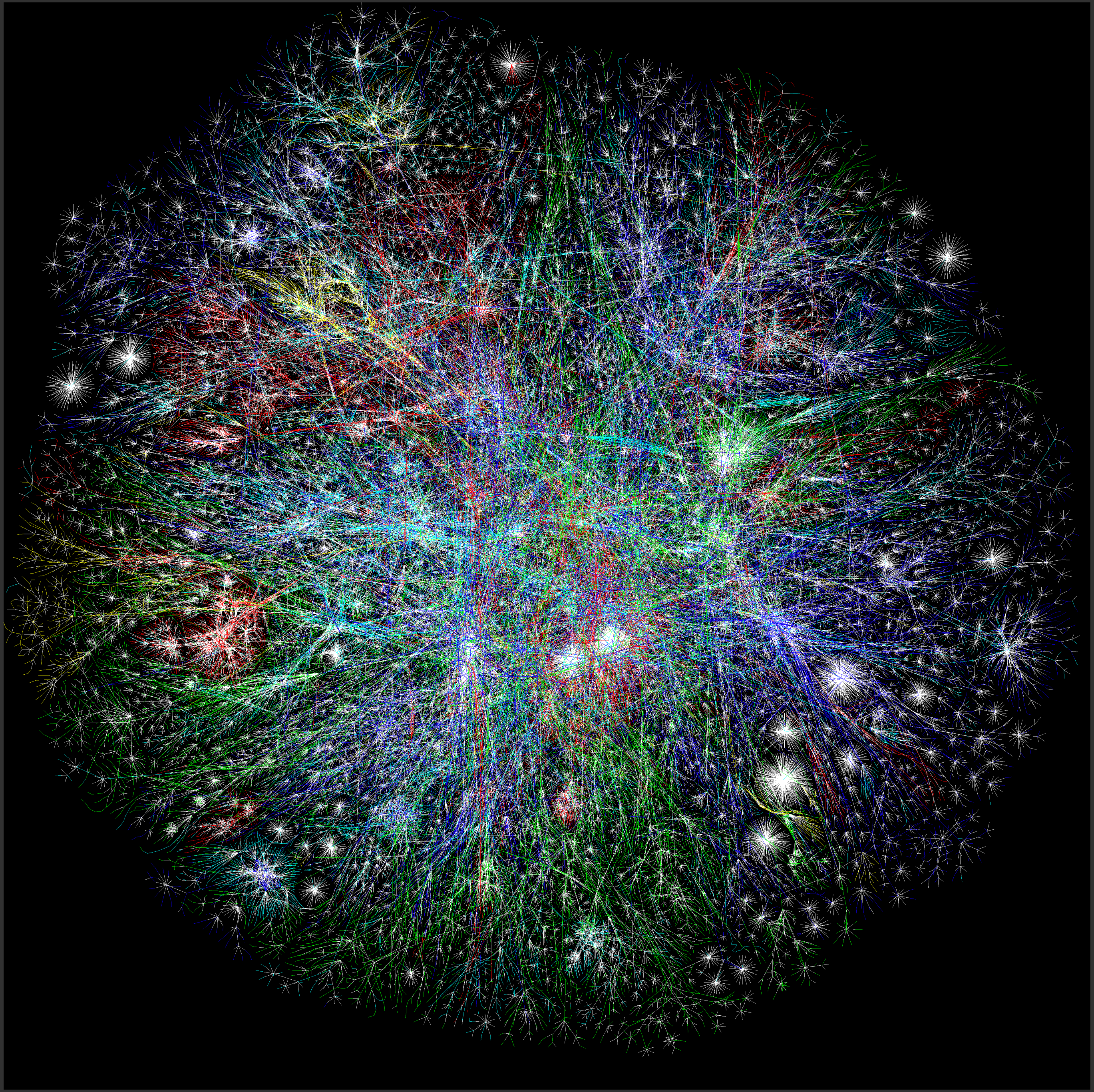
Ron Graham (alias Tom Odda).

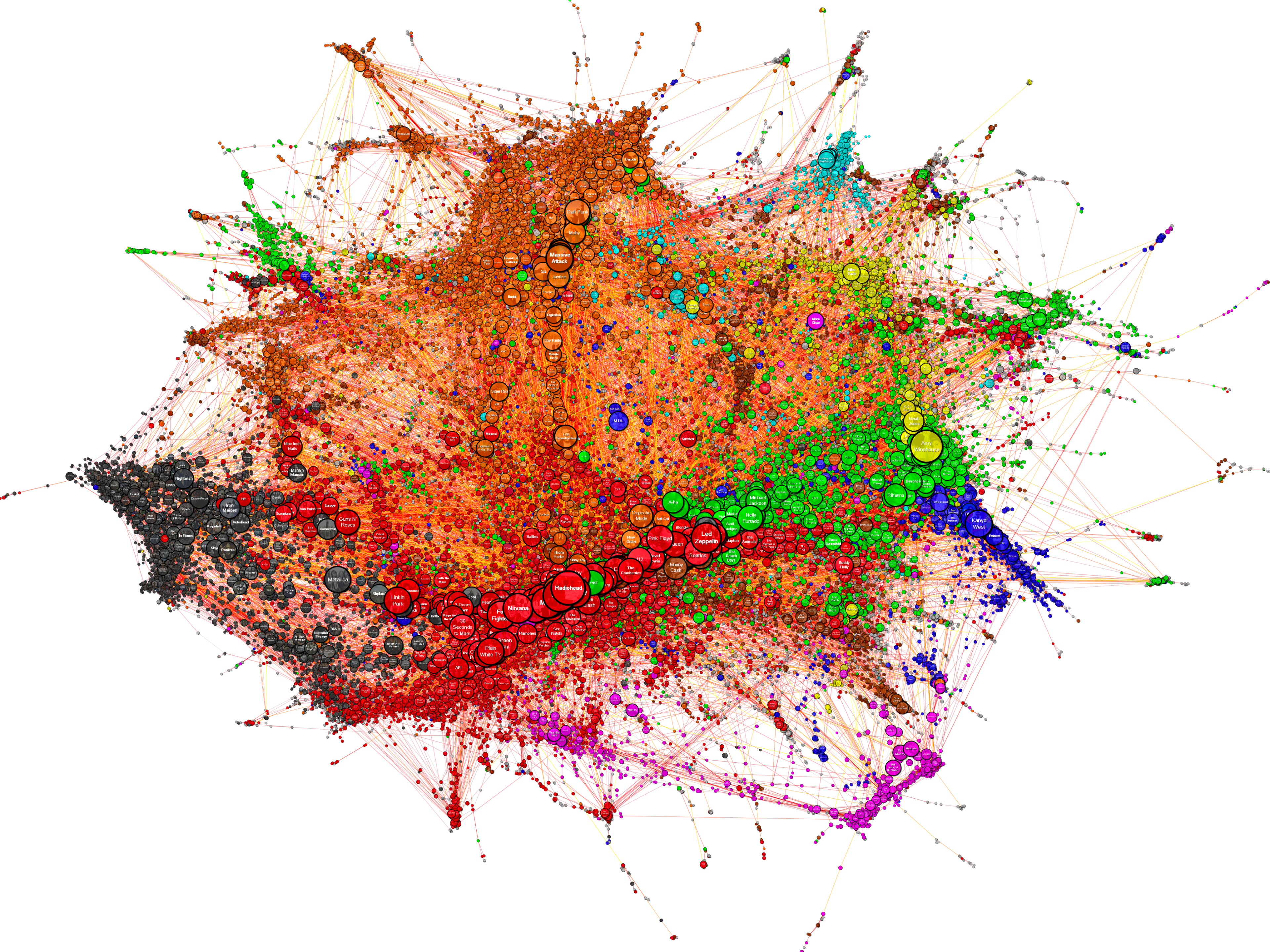
Figure 1

To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

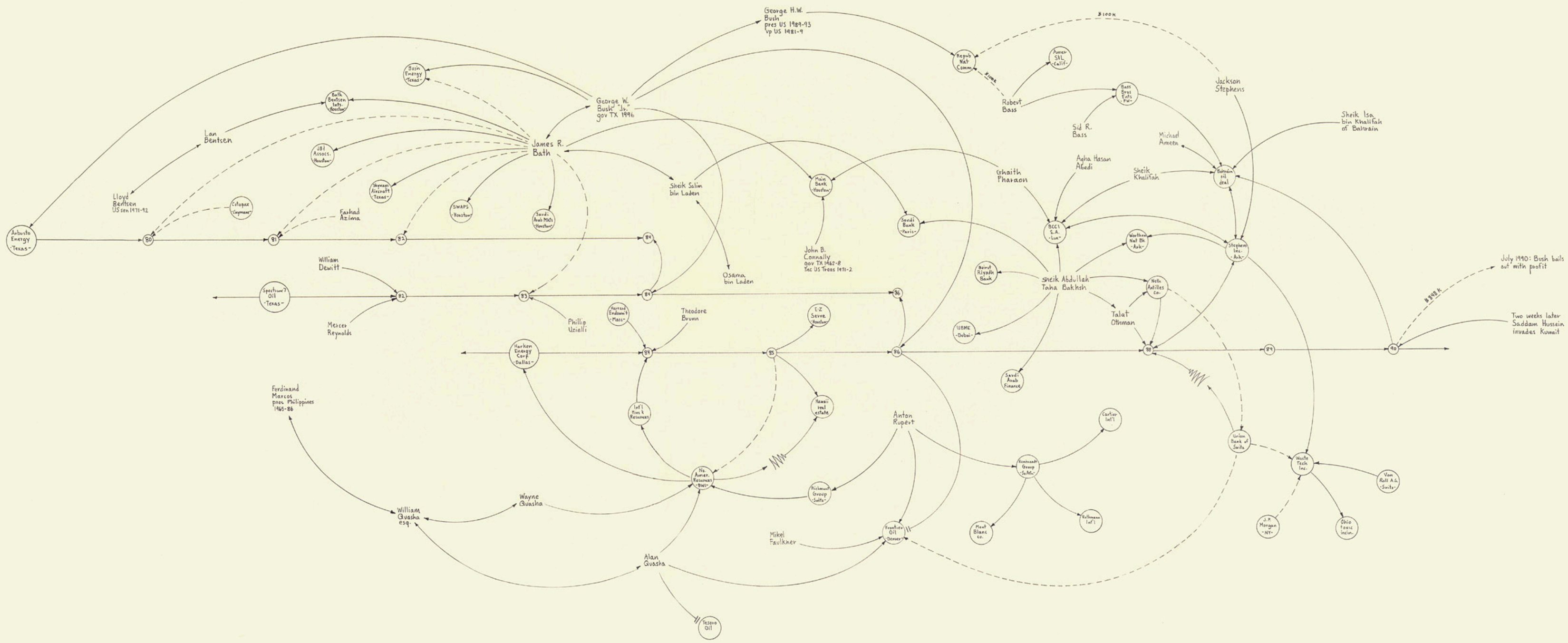


Bryan Cranston
LARRY CROWNE
Wilmer Valderrama
BEAUTY SHOP
Vivien Leigh
GONE WITH THE WIND
George Reeves
WINGED VICTORY
Kevin McCarthy
HERO AT LARGE
Arnold Schwarzenegger
ERASER
John Slattery
SLEEPERS
Ben Kingsley
PARTING SHOTS
John Cleese
Ingrid Bergman
ADAM HAD FOUR SONS
June Lockhart
MEET ME IN ST. LOUIS
Judy Garland
THE BIG PICTURE
Roddy McDowall
Eddie Albert
THE LIFE AND TIMES OF JUDGE ROY BEAN
Paul Newman
THE WAGONS ROLL AT NIGHT
Humphrey Bogart
A FEW GOOD MEN
Tom Cruise
MINORITY REPORT
Cameron Diaz
CONFESSIONS OF A DANGEROUS MIND
George Clooney
Julia Roberts
FLATLINERS
MYSTIC RIVER
Laura Linney
ABSOLUTE POWER
Clint Eastwood
Sidney Poitier
THE SLENDER THREAD
Ed Asner
STRANGERS WHEN WE MEET
Kirk Douglas
JFK
Walter Matthau
WHO'S GOT THE ACTION
Lana Turner
Laurie Metcalf
Jack Lemmon
DAD
Kevin Spacey
BULWORTH
Halle Berry
BEYOND ALL BOUNDARIES
Gary Sinise
FORREST GUMP
Sally Field
MY DOG SKIP
Diane Lane
CATTLE ANNIE AND LITTLE BRITCHES
Burt Lancaster
Emile Hirsch
INTO THE WILD
Kristen Stewart
THE AIR I BREATHE
Forest Whitaker
THE GREAT DEBATERS
Denzel Washington
Source: oracleofbacon.org

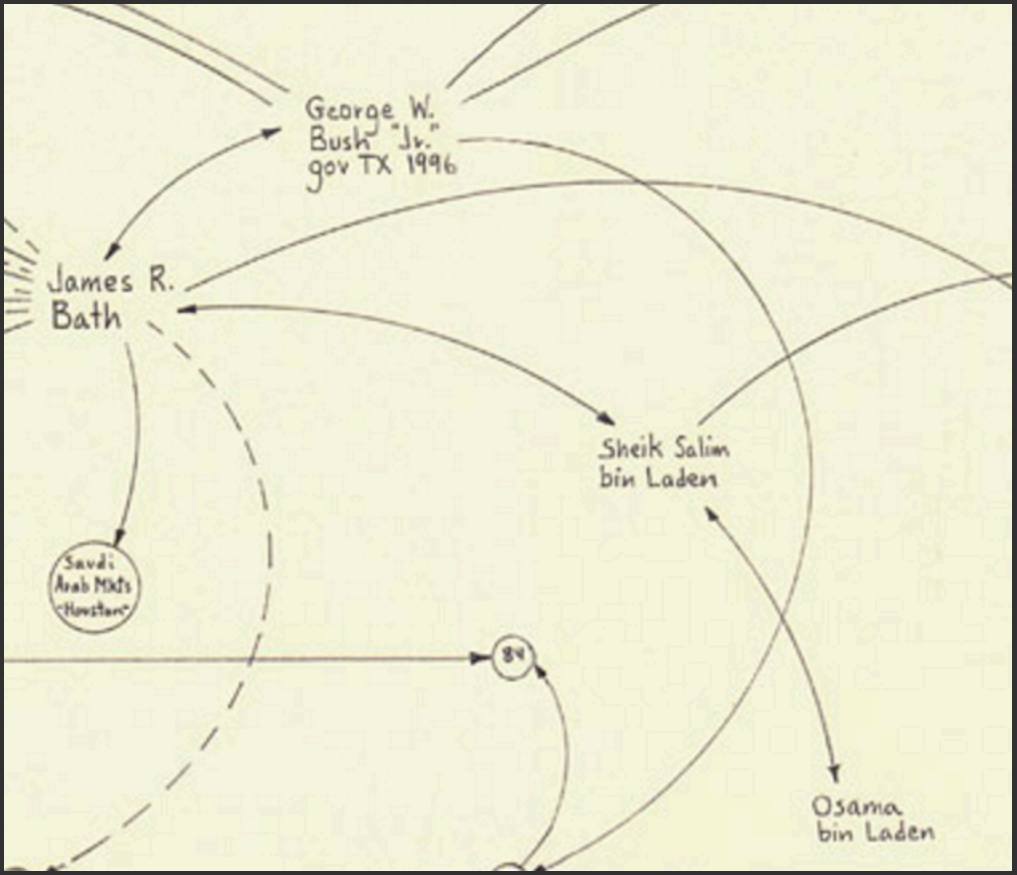




Daft Punk
Moby
Massive Attack
Justice
M.I.A.
Survivor
Manu Chao
Amy Winehouse
Kanye West
Eminem
Rihanna
Beyonce
Michael Jackson
Nelly Furtado
A-ha
Depeche Mode
New Order
Led Zeppelin
Pink Floyd
Beatles
Johnny Cash
The Cranberries
The Animals
Beach Boys
Buddy Holly
Radiohead
Nirvana
Ramones
Sex Pistols
Plain White T's
30 Seconds to Mars
AFI
Tool
Linkin Park
Slipknot
Metallica
Killswitch Engage
Nine Inch Nails
Marilyn Manson
Guns N' Roses
Europe
Van Halen
Scorpions
Rammstein
Nightwish
Iron Maiden
Megadeth
Motorhead
In Flames
Slayer
Pantera

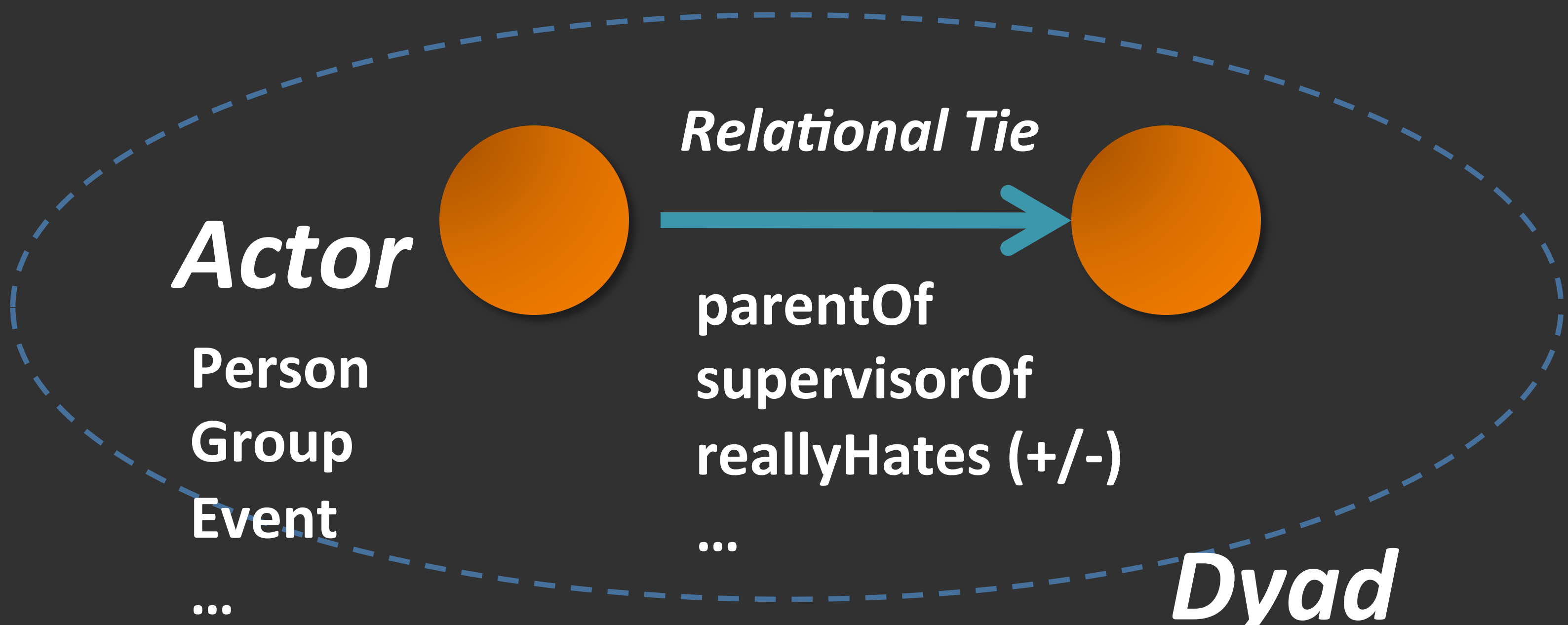


George W. Bush, Harken Energy and
Jackson Stephens c. 1979-90
5th Version
ML © 1999



George W. Bush, Harken Energy and
Jackson Stephens c. 1979-90
5th Version
ML © 1999

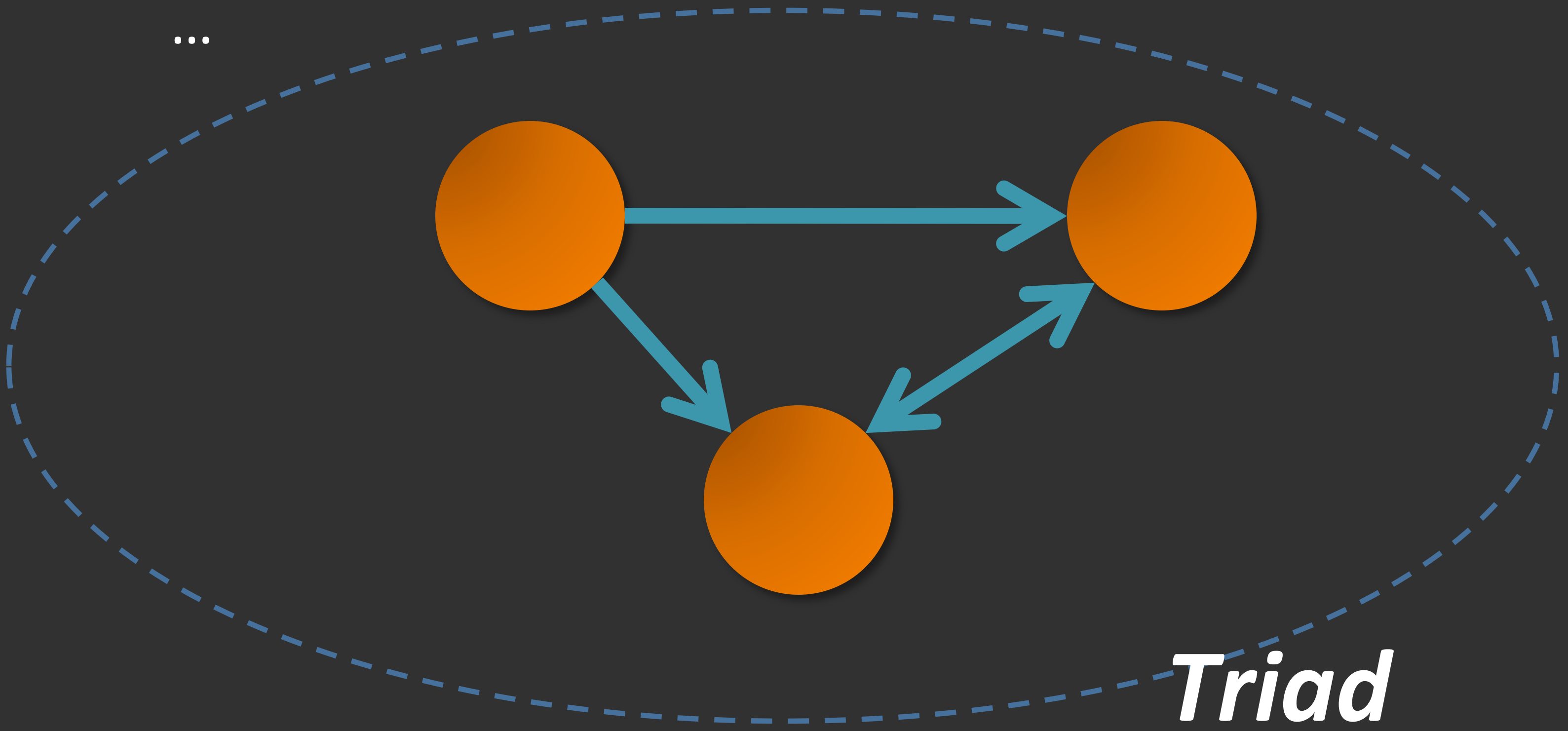
Vocabulary Lesson



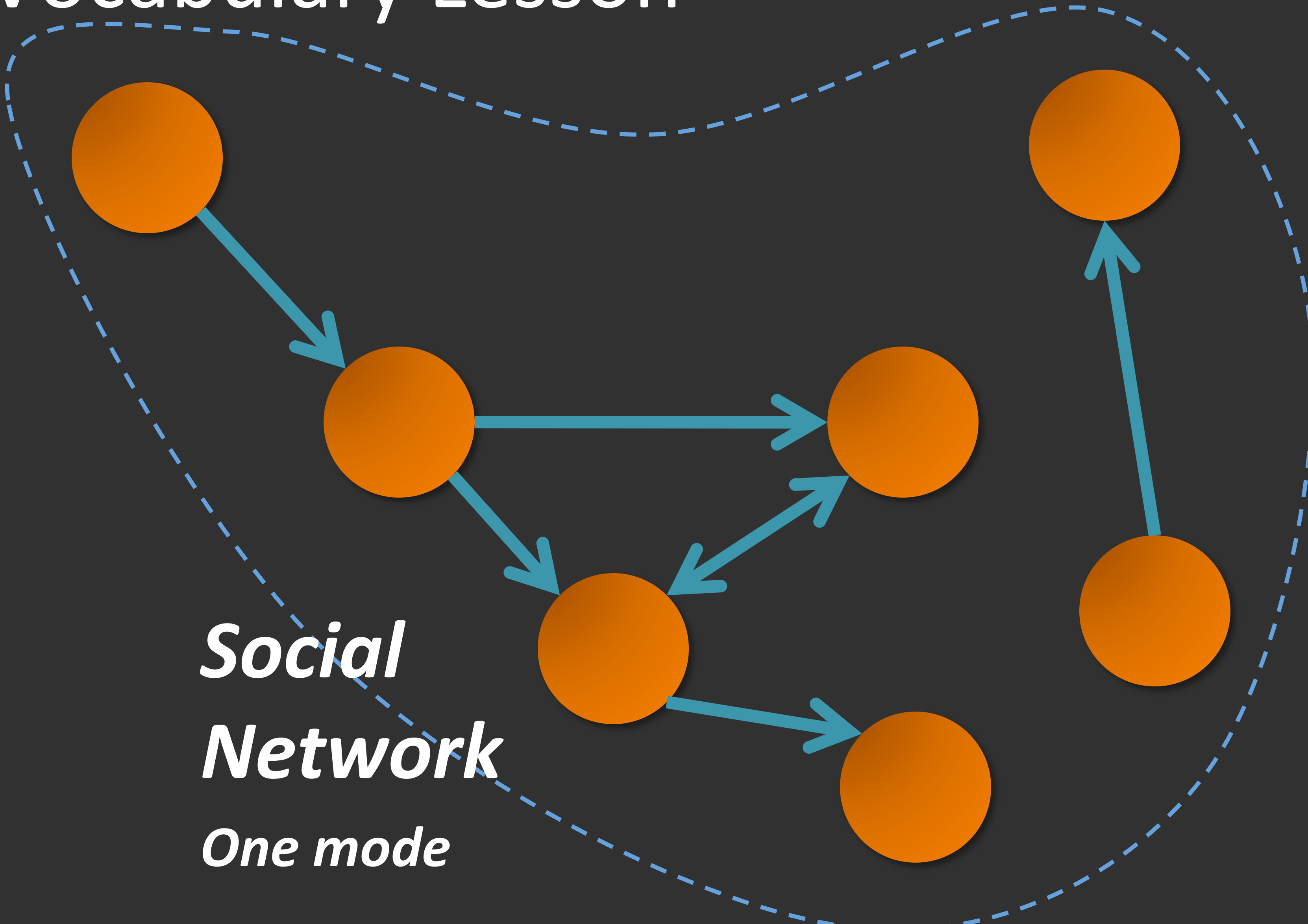
Relation: collection of ties of a specific type (every parentOf tie)

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
…



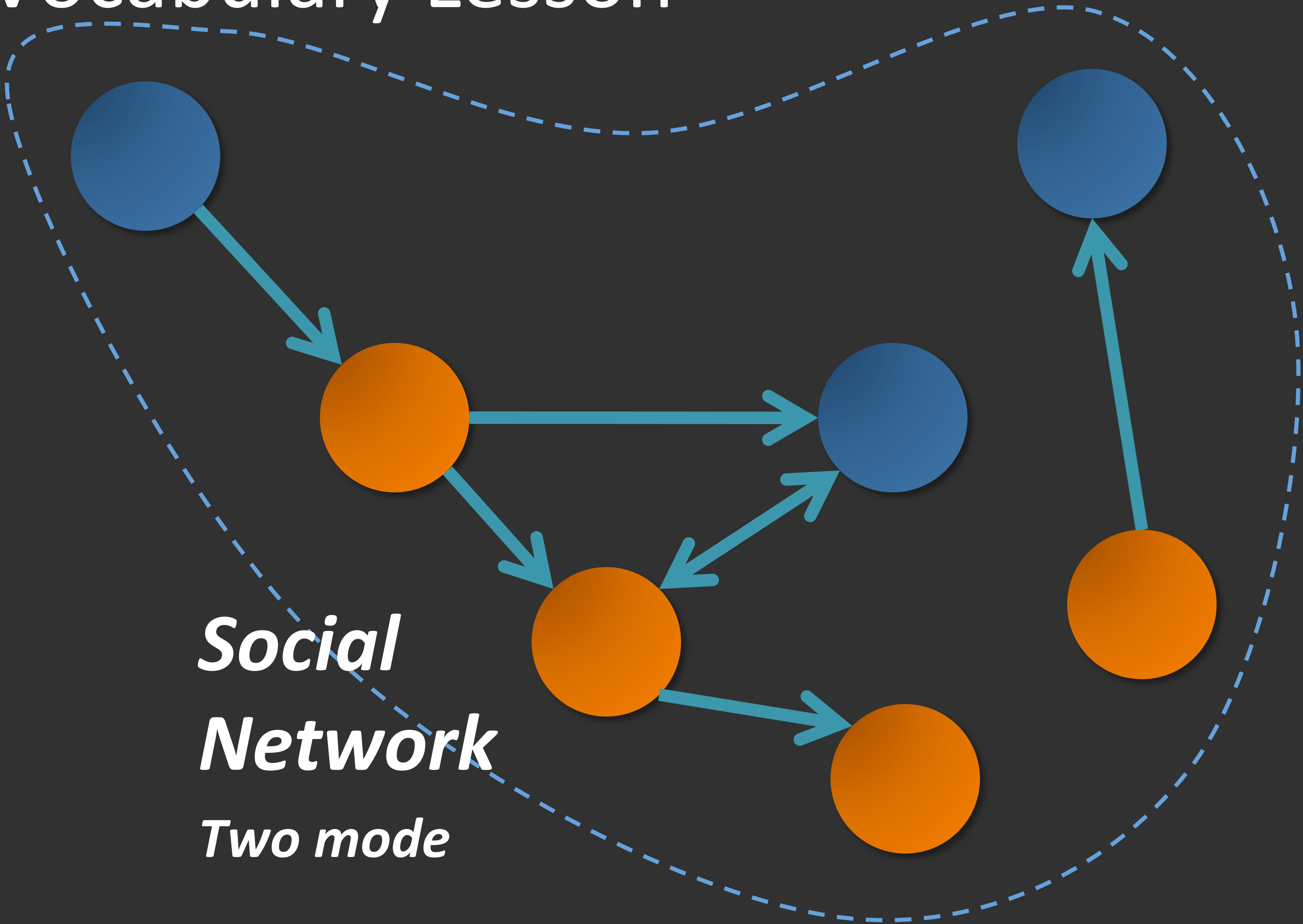
Vocabulary Lesson



Social Network

One mode

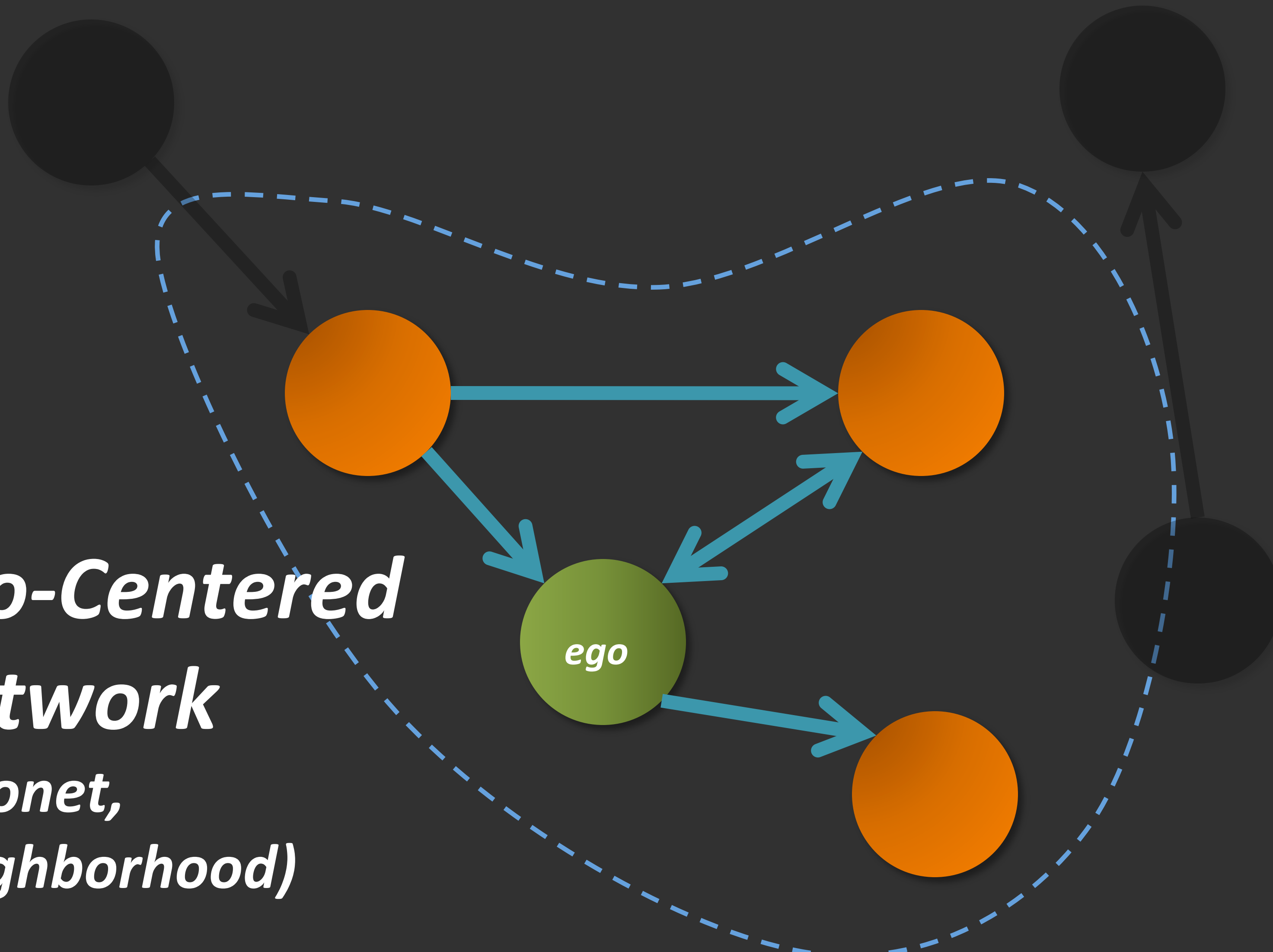
Vocabulary Lesson



Social Network

Two mode

Vocabulary Lesson



Ego-Centered Network

(egonet, neighborhood)

Describing Networks

- *Geodesic*
  - shortest_path(n,m)
- *Diameter*
  - max(geodesic(n,m)) n,m actors in graph
- *Density / Sparsity*
  - Number of existing edges / All possible edges
  - Degeneracy (number k such that every subgraph has a vertex of degree k or less)
    - Related to arboricity (number of forests that cover every edge)

Degeneracy in the Real World

graph	n	m	d
zachary [48]	34	78	4
dolphins [35]	62	159	4
power [47]	4,941	6,594	5
polbooks [28]	105	441	6
adjnoun [29]	112	425	6
football [15]	115	613	8
lesmis [25]	77	254	9
celegensneural [47]	297	1,248	9
netscience [39]	1,589	2,742	19
internet [40]	22,963	48,421	25
condmat-2005 [38]	40,421	175,693	29
polblogs [4]	1,490	16,715	36
astro-ph [38]	16,706	121,251	56

From https://arxiv.org/pdf/1006.5440

Degeneracy in the Real World

graph	n	m	d
mouse	1,455	1,636	6
worm	3,518	3,518	10
plant	1,745	3,098	12
fruitfly	7,282	24,894	12
human	9,527	31,182	12
fission-yeast	2,031	12,637	34
yeast	6,008	156,945	64

From https://arxiv.org/pdf/1006.5440

Degeneracy in the Real World

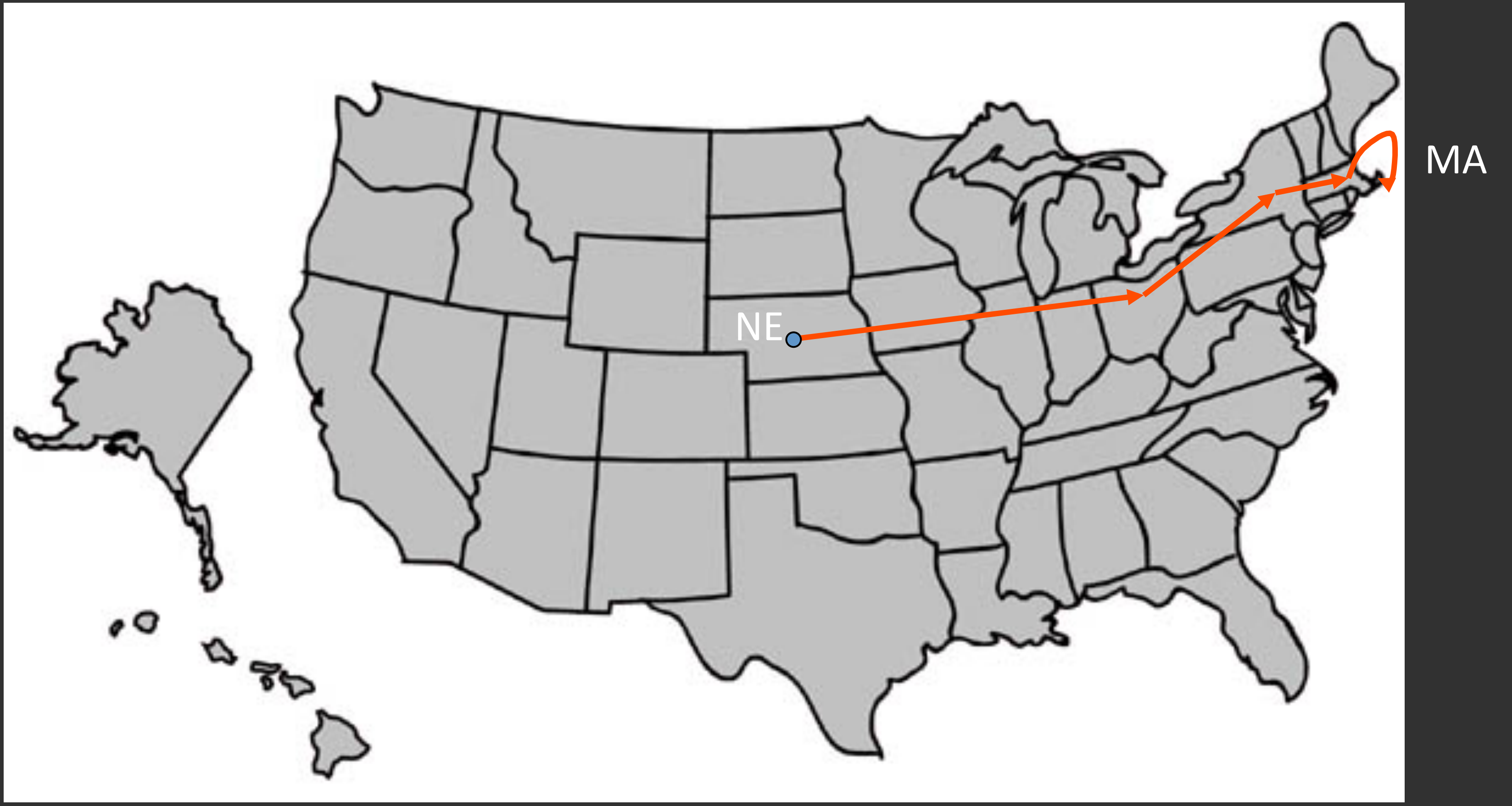
graph	n	m	d
roadNet-CA [34]	1,965,206	2,766,607	3
roadNet-PA [34]	1,088,092	1,541,898	3
roadNet-TX [34]	1,379,917	1,921,660	3
amazon0601 [30]	403,394	2,443,408	10
email-EuAll [31]	265,214	364,481	37
email-Enron [24]	36,692	183,831	43
web-Google [2]	875,713	4,322,051	44
soc-wiki-Vote [33]	7,115	100,762	53
soc-slashdot0902 [34]	82,168	504,230	55
cit-Patents [18]	3,774,768	16,518,947	64
soc-Epinions1 [42]	75,888	405,740	67
soc-wiki-Talk [33]	2,394,385	4,659,565	131
web-berkstan [34]	685,231	6,649,470	201

From https://arxiv.org/pdf/1006.5440

Random Network Graph Models

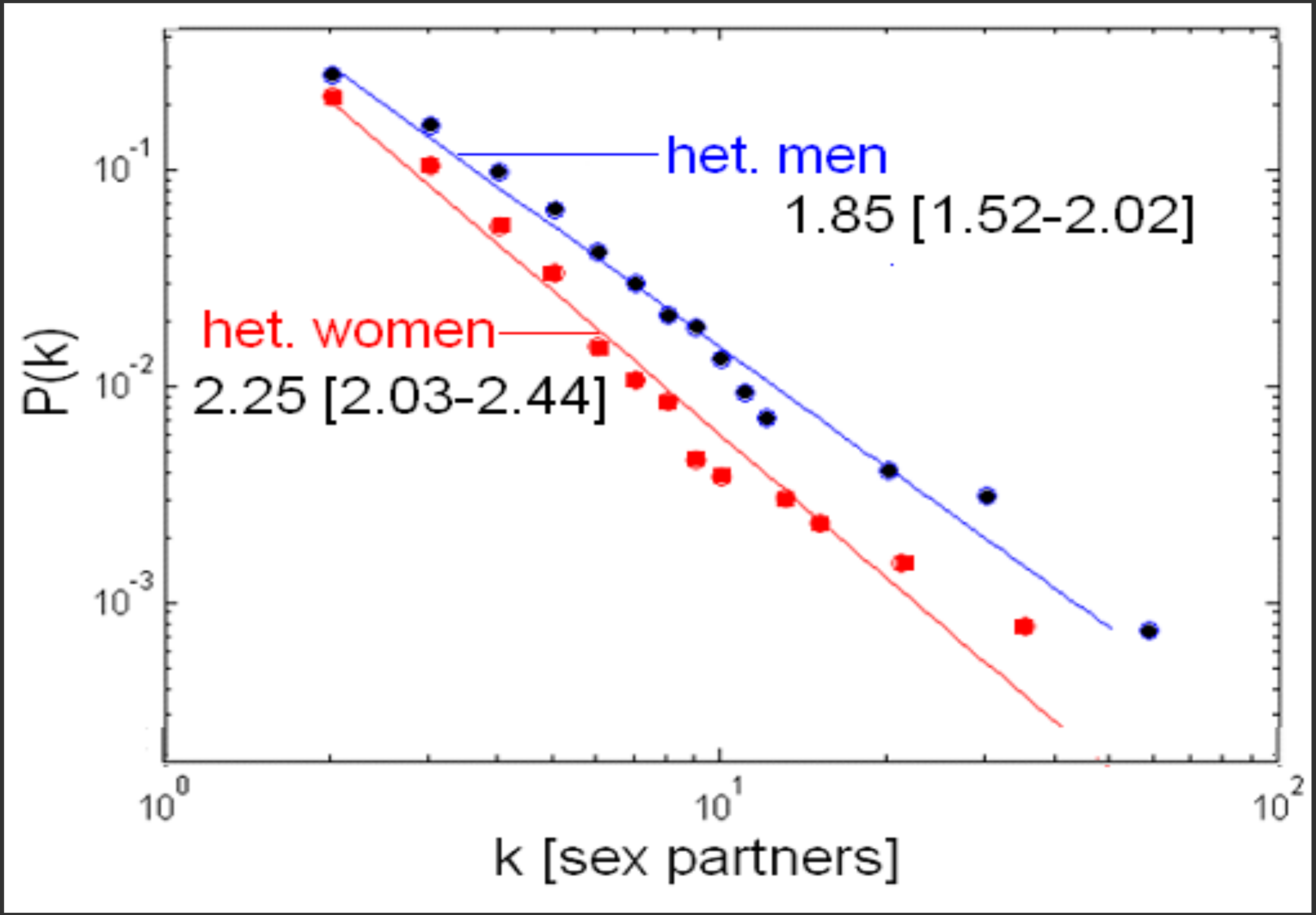
- Two classic examples:
  - Erdős–Rényi
    - *G(n,M)*: randomly draw *M* edges between *n* nodes
    - *G(n,p)*: randomly draw edges between *n* nodes, each with probability *p*.
  - These models don't really model the real world, in that they don't show:
    - Small world phenomenon
    - Power laws
    - Sparsity

Small world experiment



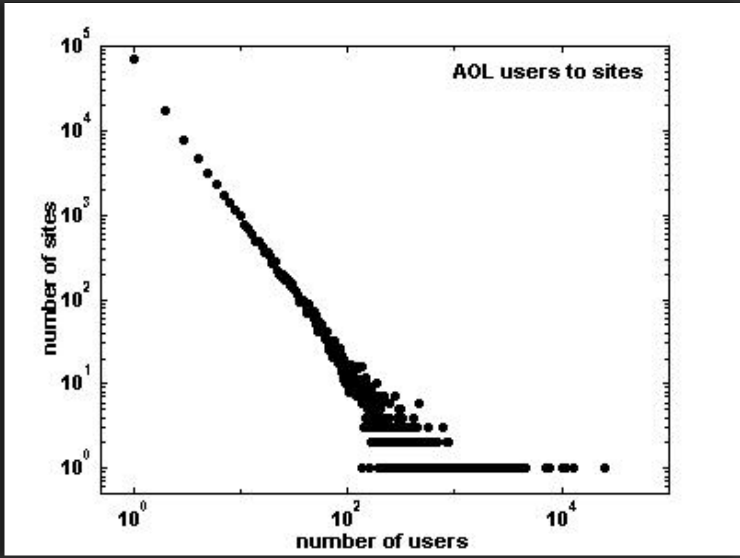
Milgram's experiment (1960's):

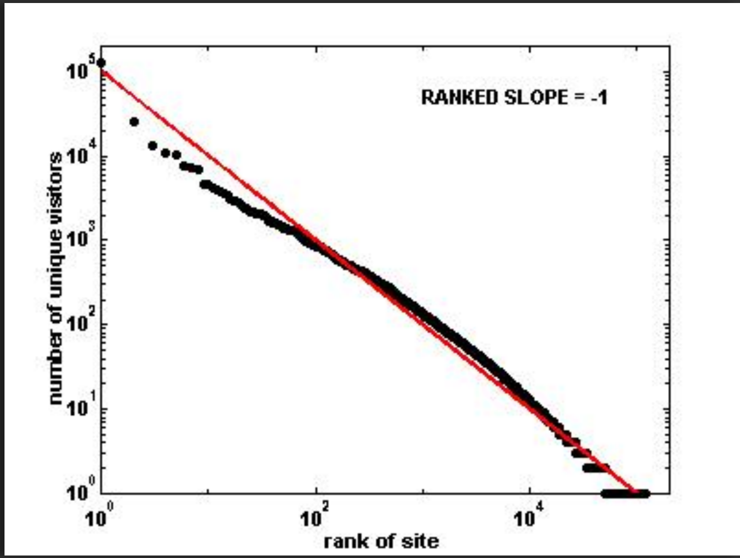
- Given a target individual and a particular property, pass the message to a person you correspond with who is "closest" to the target.
- "Six degrees of separation"


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

Two more examples of power laws

Distribution of users among web sites





Sites ranked by popularity

Power Laws (Scale-Free Networks)

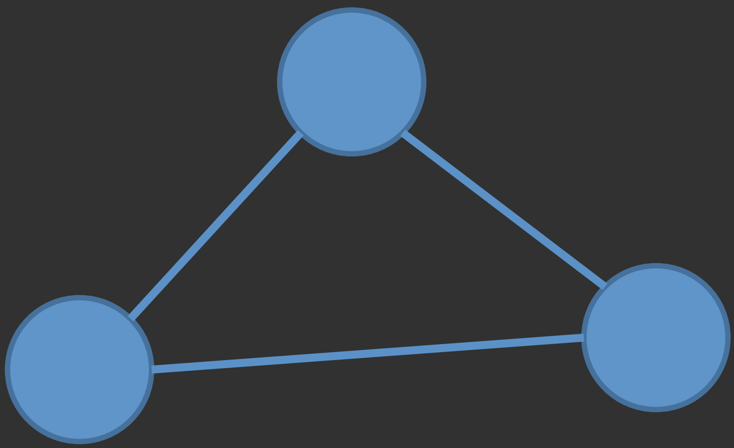
- Power-law
  - A scale-free network is a network whose degree distribution follows a power law, at least asymptotically.
  - That is, the fraction P(k) of nodes in the network having k connections to other nodes goes for large values of k as

$$P(k) \sim x^{-k}$$

  - Typically k is in the range from 2 to 3.
  - Many networks have been reported to be scale-free.

Barabási & Albert (BA) Random Graph 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 a BA graph

- The degree distribution is scale free with exponent k = 3 $P(k) = 2\ m^2/k^3$
- The graph is connected
  - Every new vertex is born with a link or several links It then connects to m 'older' vertices
  - Probability p_i of connecting to node i: $$p_i = \frac{k_i}{\sum_j k_j}$$
    - k_i is the degree of node i
- The older get 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
- Diffusion modeling
  - Epidemiological
- Clustering
  - Clustering coefficients
- Structure analysis
  - Subgraph 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)

Common Tasks

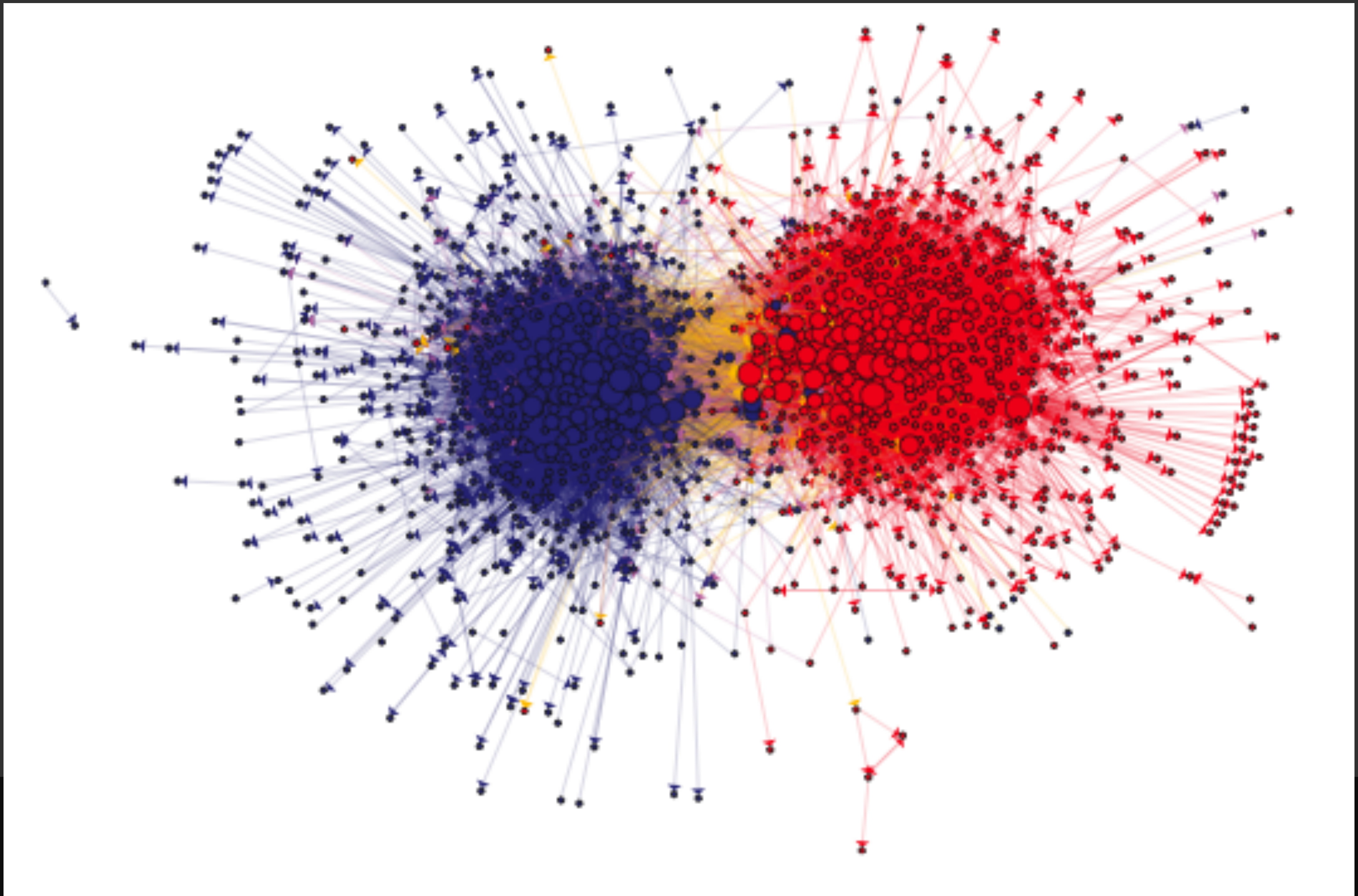
- Measuring “importance”
  - Centrality, prestige (incoming links)
- Diffusion modeling
  - Epidemiological
- Clustering
  - Clustering coefficients
- Structure analysis
  - Subgraph Isomorphisms, etc.
- Visualization/Privacy/etc.

Epidemiological

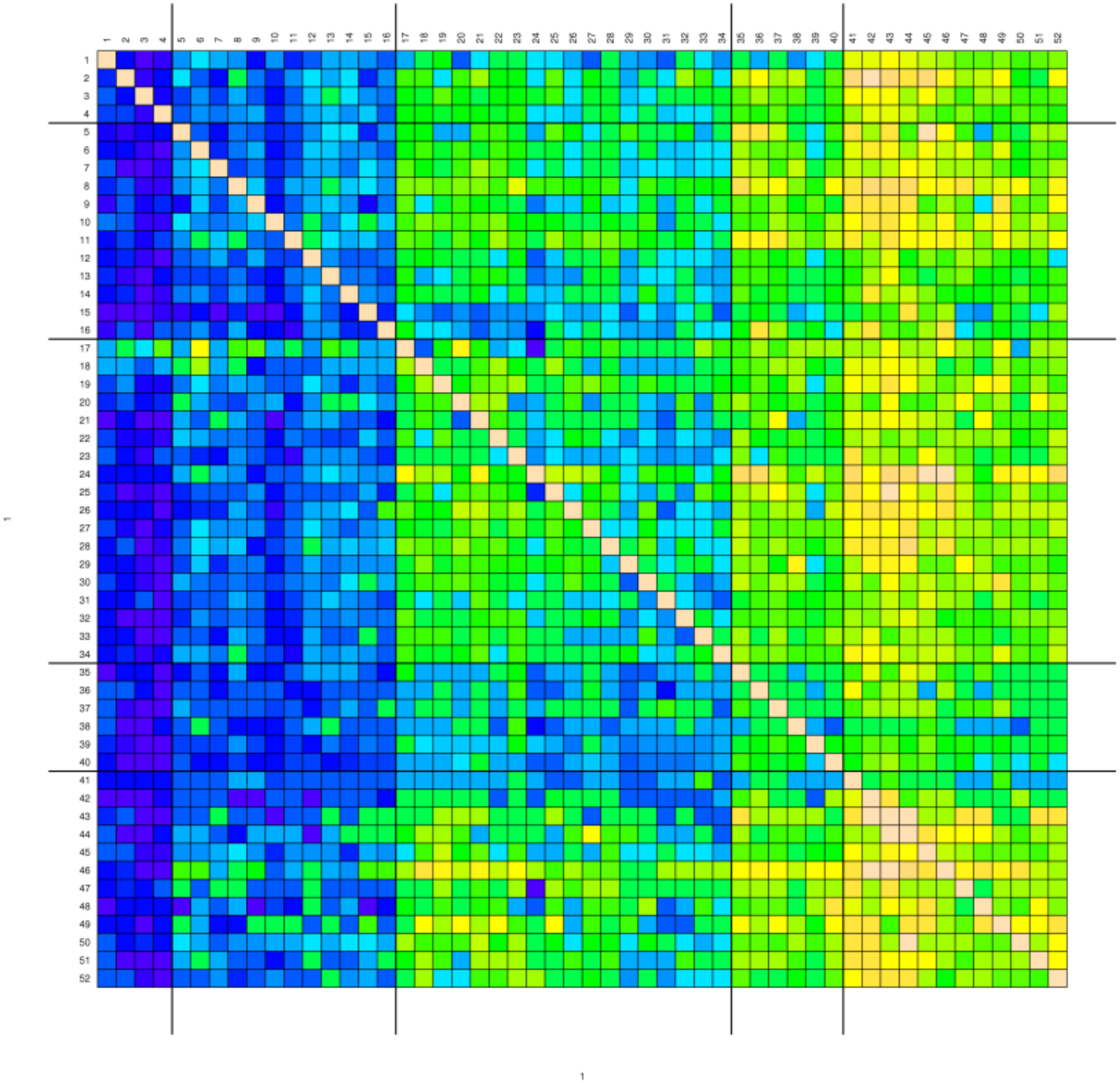
- Viruses
  - Biological, computational
  - STDs, needle sharing, etc.
  - Mark Handcock at UW
- Blog networks
  - Applying SIR models (Info Diffusion Through Blogspace, Gruhl et al.)
    - Induce transmission graph, cascade models, simulation
  - Link prediction (Tracking Information Epidemics in Blogspace,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

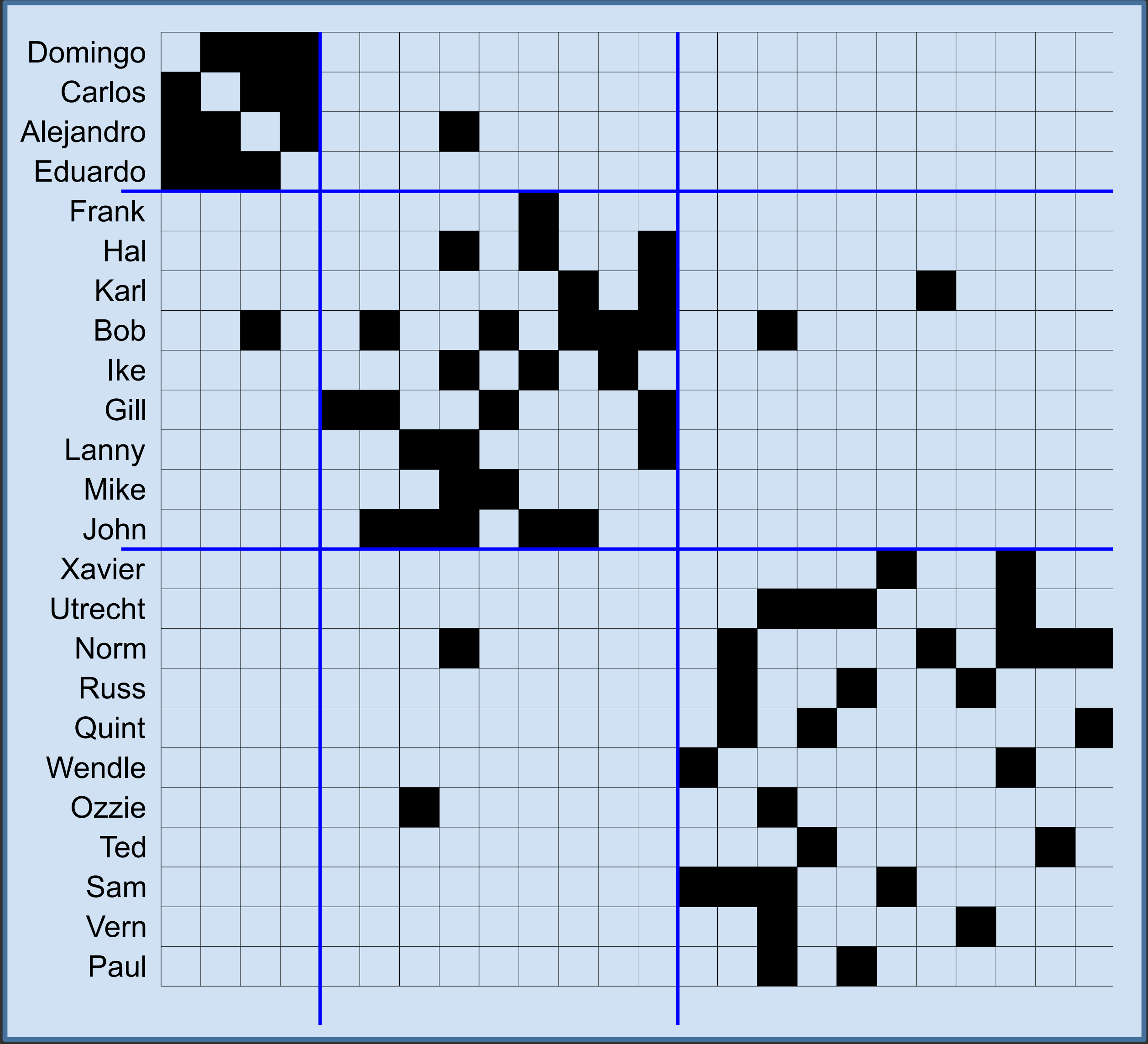
Common Tasks

- Measuring “importance”
  - Centrality, prestige (incoming links)
- Diffusion modeling
  - Epidemiological
- Clustering
  - Clustering coefficients
- Structure analysis
  - Subgraph Isomorphisms, etc.
- Visualization/Privacy/etc.



Blockmodel of U.S. Philosophy Departments. Note that row/column numbers do not correspond to PGR rankings.




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

Global Clustering Coefficient

- The global clustering coefficient C is defined as:

$$C = \frac{3 \times \text{number of triangles}}{\text{number of connected triplets of vertices}} = \frac{\text{number of closed triplets}}{\text{number of connected triplets of vertices}}.$$

- In this formula, a connected triplet is defined to be a connected subgraph consisting of three vertices and two edges. Thus, each triangle forms three connected triplets, explaining the factor of three in the formula.

Local Clustering Coefficient

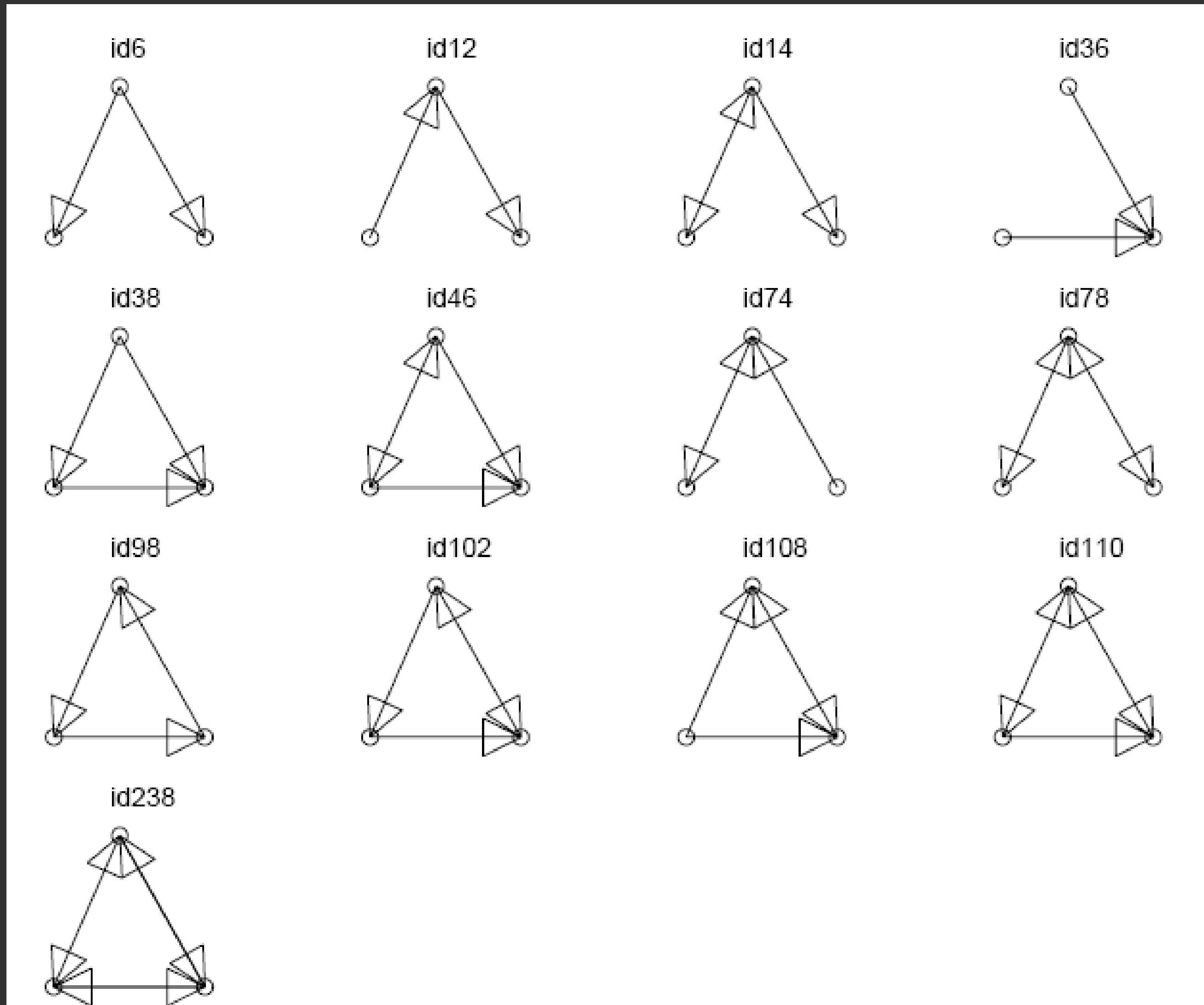
- The local clustering coefficient of a vertex (node) in a graph quantifies how close its neighbors are to being a clique (i.e., complete graph).
- The number of possible connections for the neighbors of a node i of degree k_i is, of course, $k_i(k_i-1)/2$.
- The local clustering coefficient C_i of node i is defined as:

$$C_i = \frac{2|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{k_i(k_i - 1)}.$$

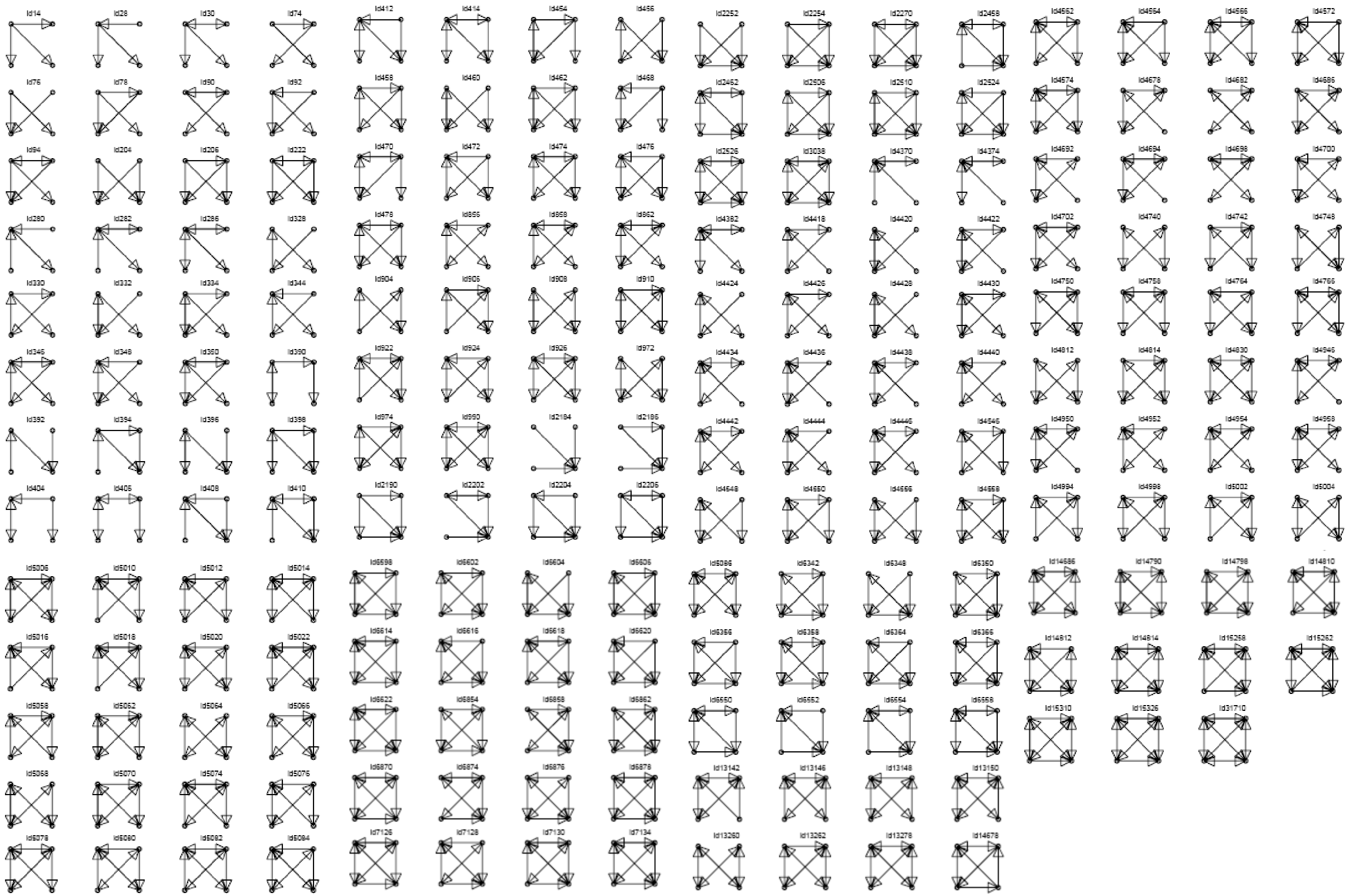
- We will discuss later how to compute these values.

Common Tasks

- Measuring “importance”
  - Centrality, prestige (incoming links)
- Diffusion modeling
  - Epidemiological
- Clustering
  - Blockmodeling, Girvan-Newman
- Structure analysis
  - Subgraph 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 Id2458 Id4562 Id4564 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 Id906 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 Id4958
Id404 Id406 Id408 Id410 Id2190 Id2202 Id2204 Id2206 Id4548 Id4550 Id4556 Id4558 Id4994 Id4998 Id5002 Id5004
Id5006 Id5010 Id5012 Id5014 Id6598 Id6602 Id6604 Id6606 Id6086 Id6342 Id6348 Id6350 Id14686 Id14790 Id14798 Id14810
Id5016 Id5018 Id5020 Id5022 Id6614 Id6616 Id6618 Id6620 Id6356 Id6358 Id6364 Id6366 Id14812 Id14814 Id15258 Id15262
Id5058 Id5062 Id5064 Id5066 Id6622 Id6854 Id6858 Id6862 Id6650 Id6652 Id6654 Id6658 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

Common Tasks

- Measuring “importance”
  - Centrality, prestige (incoming links)
- Diffusion modeling
  - Epidemiological
- Clustering
  - Clustering coefficients
- 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