# Week 12

# Agenda

Today: Network Science discussion

For next week: Read Gomez paper

**Networks or Graphs?**

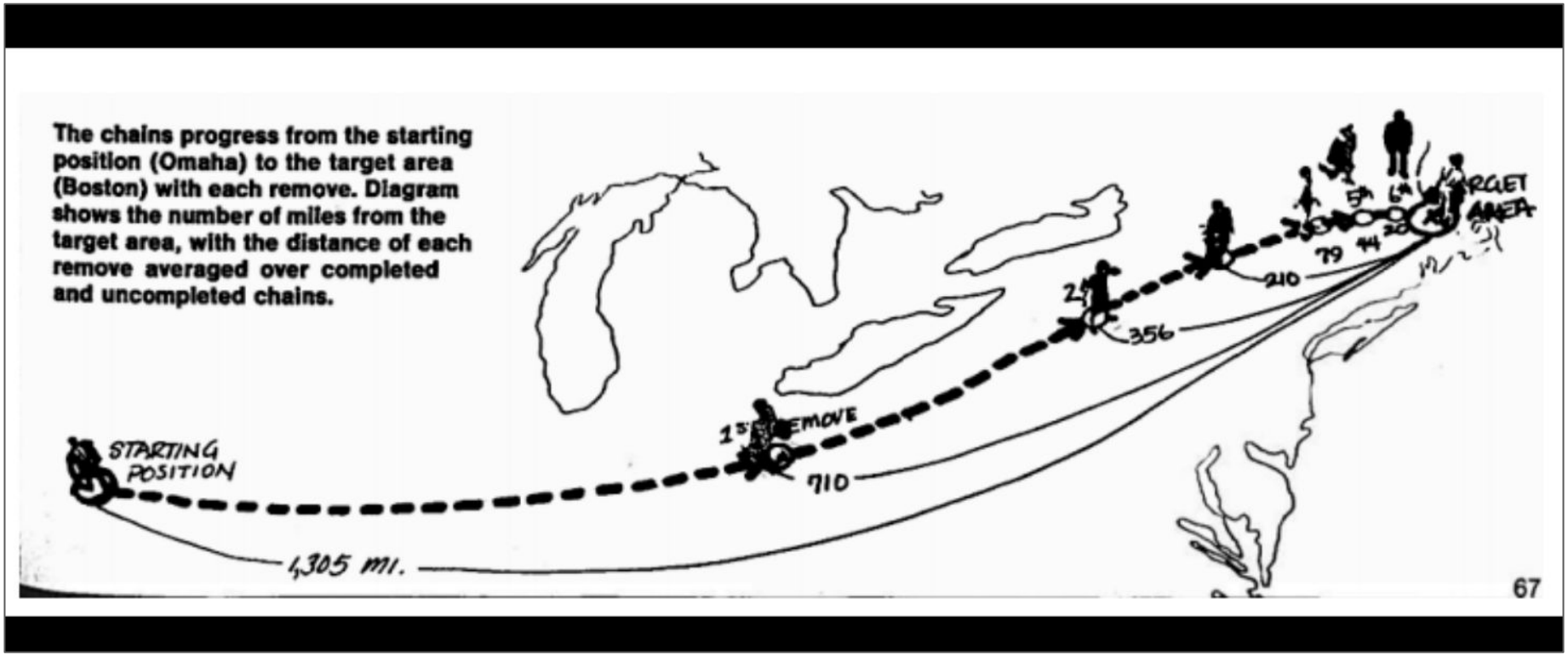
In the scientific literature the terms *network* and *graph* are used interchangeably:

| Network Science | Graph Teory |
|---|---|
| Network | Graph |
| Node | Vertex |
| Link | Edge |

Yet, there is a subtle distinction between the two terminologies: the {*network*, *node*, *link*} combination often refers to real systems: The WWW is a network of web documents linked by URLs; society is a network of individuals linked by family, friendship or professional ties; the metabolic network is the sum of all chemical reactions that take place in a cell. In contrast, we use the terms {*graph*, *vertex*, *edge*} when we discuss the mathematical representation of these networks: We talk about the web graph, the social graph (a term made popular by Facebook), or the metabolic graph. Yet, this distinction is rarely made, so these two terminologies are often synonyms of each other.

# Stanley Milgram (Visual Approximation)

# Stanley Milgram



# Previous Era

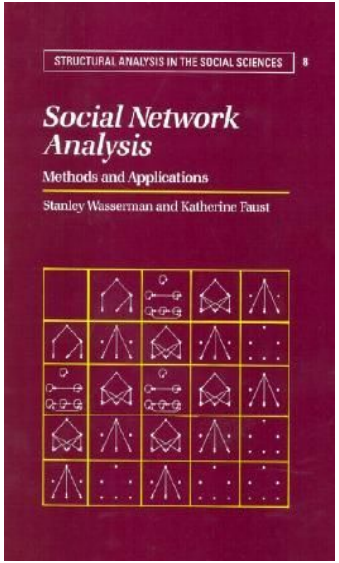



Table of Contents

Part I. Introduction: Networks, Relations, and Structure:
1. Relations and networks in the social and behavioral sciences
2. Social network data: collection and application
Part II. Mathematical Representations of Social Networks:
3. Notation
4. Graphs and matrixes
Part III. Structural and Locational Properties:
5. Centrality, prestige, and related actor and group measures
6. Structural balance, clusterability, and transitivity
7. Cohesive subgroups
8. Affiliations, co-memberships, and overlapping subgroups
Part IV. Roles and Positions:
9. Structural equivalence
10. Blockmodels
11. Relational algebras
12. Network positions and roles
Part V. Dyadic and Triadic Methods:
13. Dyads
14. Triads
Part VI. Statistical Dyadic Interaction Models:
15. Statistical analysis of single relational networks
16. Stochastic blockmodels and goodness-of-fit indices
Part VII. Epilogue:
17. Future directions.

In the previous era, it was difficult to gather or uncover data.

**1950's Random Graphs** (Paul Erdos)

**1960's Small world experiment** (Stanley Milgram)

- http://en.wikipedia.org/wiki/Milgram_experiment.
- Many letters never arrived.
- Average path length is 5.5-6.
- Important people: "mr jacobs".
- FB has avg path length of 4.75, twitter 4.67, MS Messanger 6.6
- Tended to arrive at Geography, but difficult path to individual. (Idea of 5 is traced to 1929)

**1990's Wasserman**

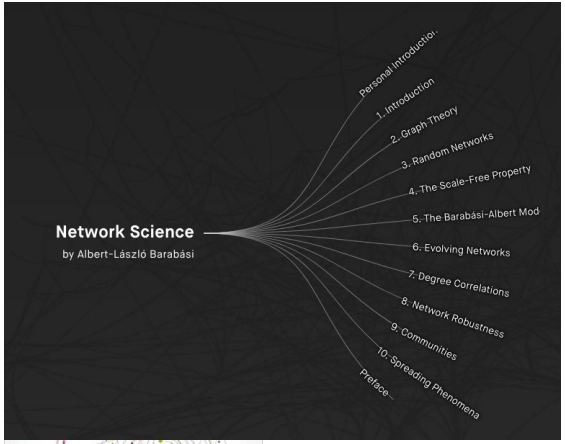
- http://www.amazon.com/Social-Network-Analysis-Applications-Structural/dp/0521387078/x

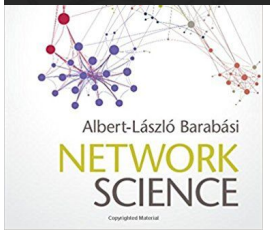
# Recent History

TABLE II. The scaling exponents characterizing the degree distribution of several scale-free networks, for which $P(k)$ follows a power-law (2). We indicate the size of the network, its average degree $\langle k \rangle$ and the cutoff $\kappa$ for the power-law scaling. For directed networks we list separately the indegree ($\gamma_{in}$) and outdegree ($\gamma_{out}$) exponents, while for the undirected networks, marked with a star, these values are identical. The columns $l_{real}$, $l_{rand}$ and $l_{pow}$ compare the average path length of real networks with power-law degree distribution and the prediction of random graph theory (17) and that of Newman, Strogatz and Watts (2000) (62), as discussed in Sect. V. The last column identifies the symbols in Figs. 8 and 9.

| Network | Size | $\langle k \rangle$ | $\kappa$ | $\gamma_{out}$ | $\gamma_{in}$ | $\ell_{real}$ | $\ell_{rand}$ | $\ell_{pow}$ | Reference | Nr. |
|---|---|---|---|---|---|---|---|---|---|---|
| WWW | 325,729 | 4.51 | 900 | 2.45 | 2.1 | 11.2 | 8.32 | 4.77 | Albert, Jeong, Barabási 1999 | 1 |
| WWW | $4 \times 10^7$ | 7 | | 2.38 | 2.1 | | | | Kumar *et al.* 1999 | 2 |
| WWW | $2 \times 10^8$ | 7.5 | 4,000 | 2.72 | 2.1 | 16 | 8.85 | 7.61 | Broder *et al.* 2000 | 3 |
| WWW, site | 260,000 | | | | 1.94 | | | | Huberman, Adamic 2000 | 4 |
| Internet, domain* | 3,015 - 4,389 | 3.42 - 3.76 | 30 – 40 | 2.1 - 2.2 | 2.1 - 2.2 | 4 | 6.3 | 5.2 | Faloutsos 1999 | 5 |
| Internet, router* | 3,888 | 2.57 | 30 | 2.48 | 2.48 | 12.15 | 8.75 | 7.67 | Faloutsos 1999 | 6 |
| Internet, router* | 150,000 | 2.66 | 60 | 2.4 | 2.4 | 11 | 12.8 | 7.47 | Govindan 2000 | 7 |
| Movie actors* | 212,250 | 28.78 | 900 | 2.3 | 2.3 | 4.54 | 3.65 | 4.01 | Barabási, Albert 1999 | 8 |
| Coauthors, SPIRES* | 56,627 | 173 | 1,100 | 1.2 | 1.2 | 4 | 2.12 | 1.95 | Newman 2001b,c | 9 |
| Coauthors, neuro.* | 209,293 | 11.54 | 400 | 2.1 | 2.1 | 6 | 5.01 | 3.86 | Barabási *et al.* 2001 | 10 |
| Coauthors, math* | 70,975 | 3.9 | 120 | 2.5 | 2.5 | 9.5 | 8.2 | 6.53 | Barabási *et al.* 2001 | 11 |
| Sexual contacts* | 2810 | | | 3.4 | 3.4 | | | | Liljeros *et al.* 2001 | 12 |
| Metabolic, E. coli | 778 | 7.4 | 110 | 2.2 | 2.2 | 3.2 | 3.32 | 2.89 | Jeong *et al.* 2000 | 13 |
| Protein, S. cerev.* | 1870 | 2.39 | | 2.4 | 2.4 | | | | Mason *et al.* 2000 | 14 |
| Ythan estuary* | 134 | 8.7 | 35 | 1.05 | 1.05 | 2.43 | 2.26 | 1.71 | Montoya, Solé 2000 | 14 |
| Silwood park* | 154 | 4.75 | 27 | 1.13 | 1.13 | 3.4 | 3.23 | 2 | Montoya, Solé 2000 | 16 |
| Citation | 783,339 | 8.57 | | | 3 | | | | Redner 1998 | 17 |
| Phone-call | $53 \times 10^6$ | 3.16 | | 2.1 | 2.1 | | | | Aiello *et al.* 2000 | 18 |
| Words, cooccurence* | 460,902 | 70.13 | | 2.7 | 2.7 | | | | Cancho, Solé 2001 | 19 |
| Words, synonyms* | 22,311 | 13.48 | | 2.8 | 2.8 | | | | Yook *et al.* 2001 | 20 |

# Recent History

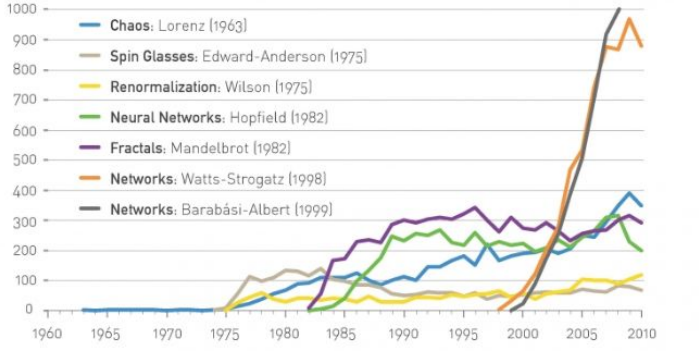



networksciencebook.com

**2000's  Scale-free (follows power law)**

- Long tail.  Fault tolerant (random attacks). Hubs.
- Examples: www, travel/airlines (and disease), protein interactions, social networks
- Scale-free: http://arxiv.org/pdf/cond-mat/0106096.pdf
- Attacks: http://www.nature.com/nature/journal/v406/n6794/abs/406378A0.html
- Barabasi book: http://www.amazon.com/Linked-Everything-Connected-Business-Everyday/dp/0465085733



# Recent History


**Today**

- Largest conference: NetSci (http://www.netsci2018.com/)
- Many research labs, both academic and industry

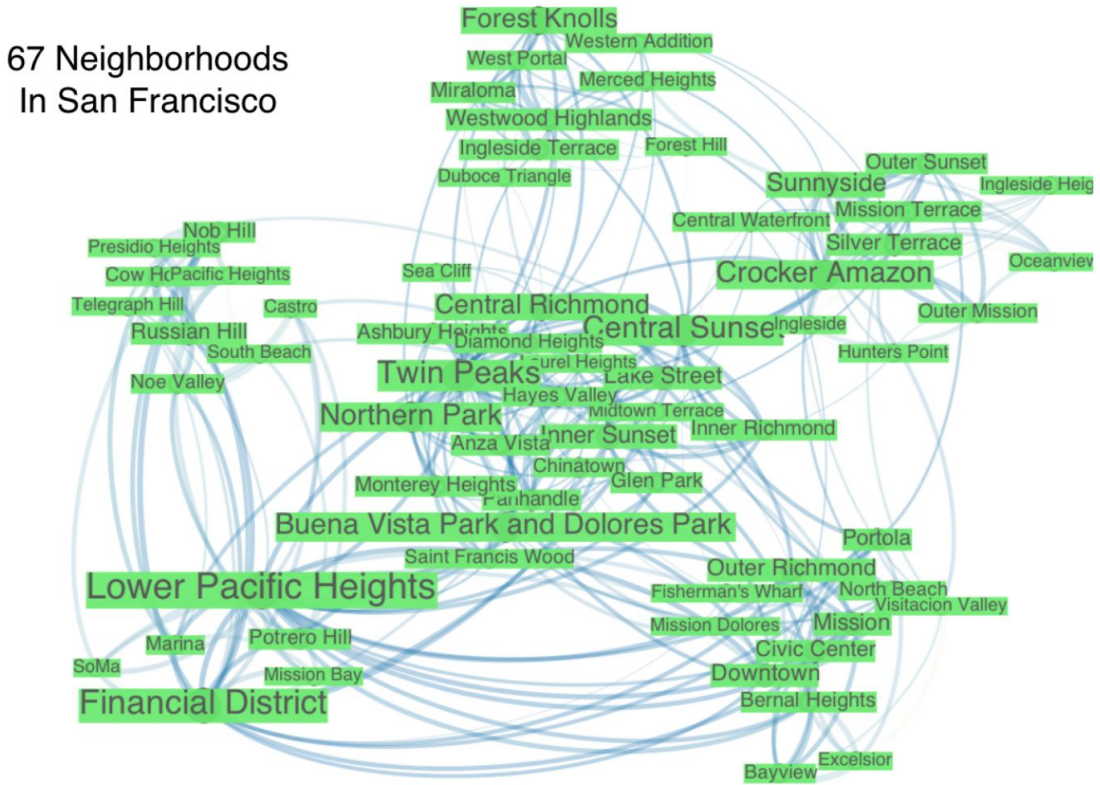
# Trulia

**Task 1**
Determine a global  ranking of neighborhoods

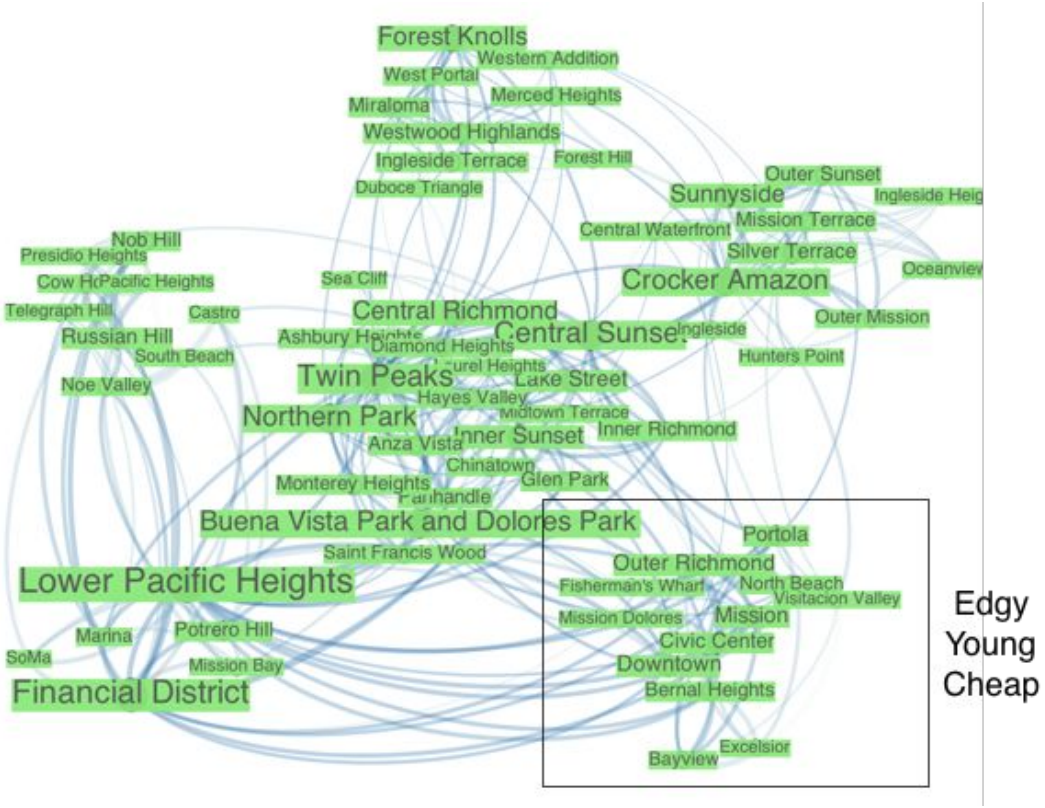
**Task 2**
Determine the relative preference of neighborhoods

**Data**
60,000 neighborhoods in US
~1B searches / clicks

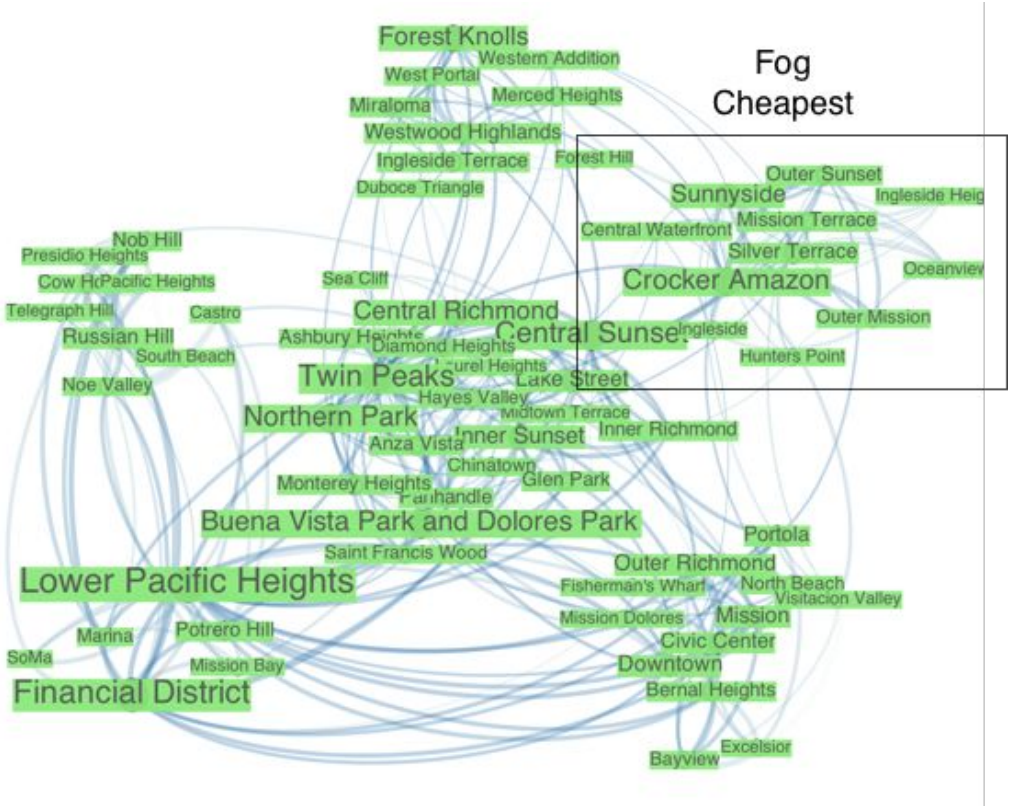
**Approach**
Construct preference graph
Task 1: Compute centrality
Task 2: Compute clusters (edge cutting)



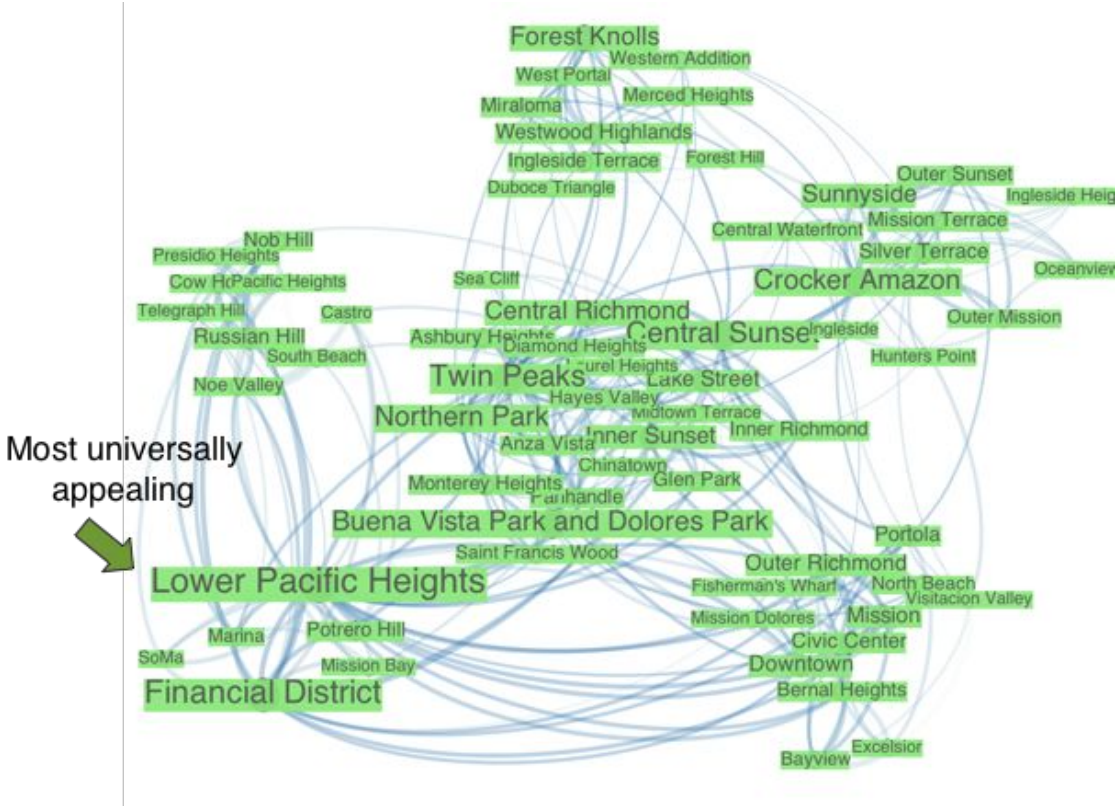
# Trulia


Forest Knolls
Western Addition
West Portal
Merced Heights
Miraloma
Westwood Highlands
Ingleside Terrace
Forest Hill
Duboce Triangle
Outer Sunset
Sunnyside
Ingleside Heig
Mission Terrace
Central Waterfront
Silver Terrace
Nob Hill
Presidio Heights
Cow HoPacific Heights
Sea Cliff
Crocker Amazon
Oceanview
Telegraph Hill
Castro
Central Richmond
Outer Mission
Russian Hill
Ashbury Heights
Central Sunset
Ingleside
Diamond Heights
South Beach
Hunters Point
Noe Valley
Twin Peaks
Lake Street
Hayes Valley
Northern Park
Midtown Terrace
Inner Sunset
Inner Richmond
Anza Vista
Chinatown
Monterey Heights
Glen Park
Buena Vista Park and Dolores Park
Portola
Saint Francis Wood
Outer Richmond
Lower Pacific Heights
Fisherman's Wharf
North Beach
Visitacion Valley
Mission Dolores
Mission
Marina
Potrero Hill
Civic Center
SoMa
Mission Bay
Downtown
Financial District
Bernal Heights
Excelsior
Bayview
Edgy
Young
Cheap

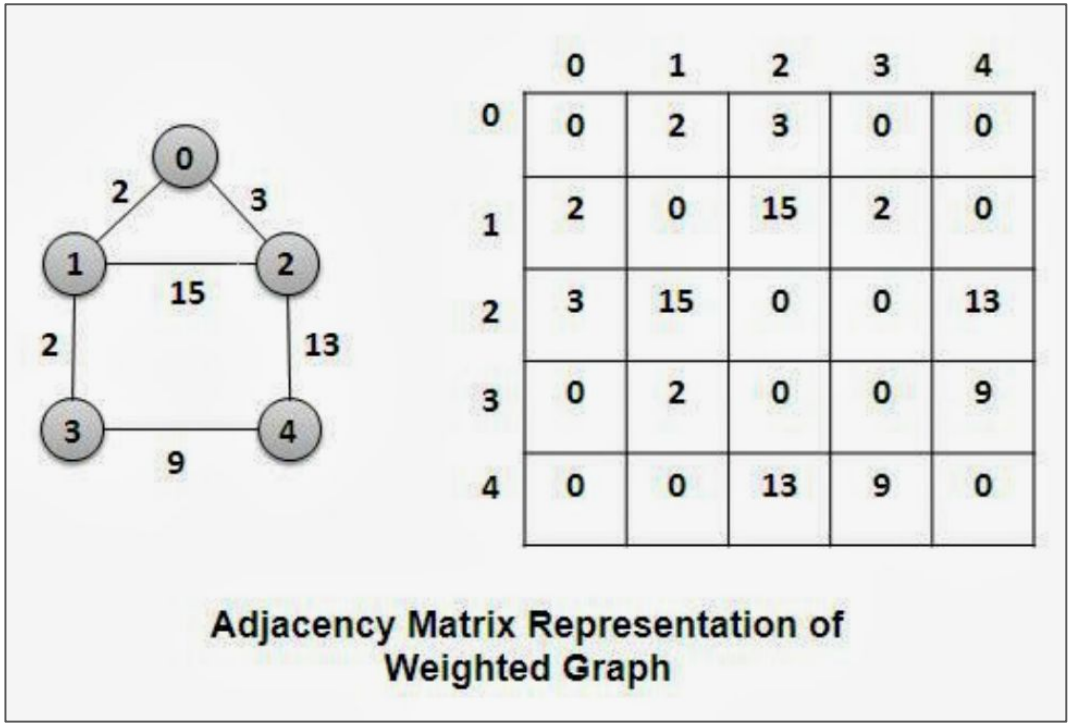
# Trulia


Forest Knolls
Western Addition
West Portal
Merced Heights
Miraloma
Westwood Highlands
Ingleside Terrace
Forest Hill
Duboce Triangle
Fog
Cheapest
Outer Sunset
Sunnyside
Ingleside Heig
Central Waterfront
Mission Terrace
Silver Terrace
Oceanview
Crocker Amazon
Nob Hill
Presidio Heights
Cow Hollow
Pacific Heights
Sea Cliff
Telegraph Hill
Castro
Central Richmond
Outer Mission
Russian Hill
Ashbury Heights
Diamond Heights
Central Sunset
Ingleside
South Beach
Laurel Heights
Hunters Point
Noe Valley
Twin Peaks
Lake Street
Hayes Valley
Midtown Terrace
Northern Park
Inner Sunset
Inner Richmond
Anza Vista
Chinatown
Monterey Heights
Glen Park
Panhandle
Buena Vista Park and Dolores Park
Portola
Saint Francis Wood
Outer Richmond
Lower Pacific Heights
Fisherman's Wharf
North Beach
Visitacion Valley
Mission Dolores
Mission
Marina
Potrero Hill
Civic Center
SoMa
Mission Bay
Downtown
Financial District
Bernal Heights
Excelsior
Bayview

# Trulia


Most universally appealing
Forest Knolls
Western Addition
West Portal
Merced Heights
Miraloma
Westwood Highlands
Ingleside Terrace
Forest Hill
Duboce Triangle
Outer Sunset
Sunnyside
Ingleside Heig
Mission Terrace
Central Waterfront
Silver Terrace
Nob Hill
Presidio Heights
Cow Hollow
Pacific Heights
Sea Cliff
Crocker Amazon
Oceanview
Telegraph Hill
Castro
Central Richmond
Outer Mission
Russian Hill
Ashbury Heights
Central Sunset
Ingleside
Diamond Heights
South Beach
Hunters Point
Noe Valley
Twin Peaks
Laurel Heights
Lake Street
Hayes Valley
Northern Park
Midtown Terrace
Anza Vista
Inner Sunset
Inner Richmond
Chinatown
Monterey Heights
Glen Park
Parnhandle
Buena Vista Park and Dolores Park
Portola
Saint Francis Wood
Outer Richmond
Lower Pacific Heights
Fisherman's Wharf
North Beach
Visitacion Valley
Mission Dolores
Mission
Marina
Potrero Hill
Civic Center
SoMa
Mission Bay
Downtown
Financial District
Bernal Heights
Excelsior
Bayview

# Dense Representation




**Dense Representation**: Adjacency matrix

- Vertex-Vertex
- 0/1 or weight

**Sparse Representation**

- Version 1: V1:[V2,V3]
- Version 2: V1:V2:nonzero-weight:node1size:node1color

**Coding Notes**

- Most NetSci libraries have separate representation of data network and visual network
- D3 is an exception

How might a sparse graph be represented?

# Statistics & Algos

**General Language**

- Graph theory vs Network Science (not hard and fast)
- Vertices/nodes (N), edges (E)
- Path (L), hops, cycles
- Undirected, directed, DAG, Bipartite
- Connectivity, cuts, component
- Hypergraphs

**Node Level**

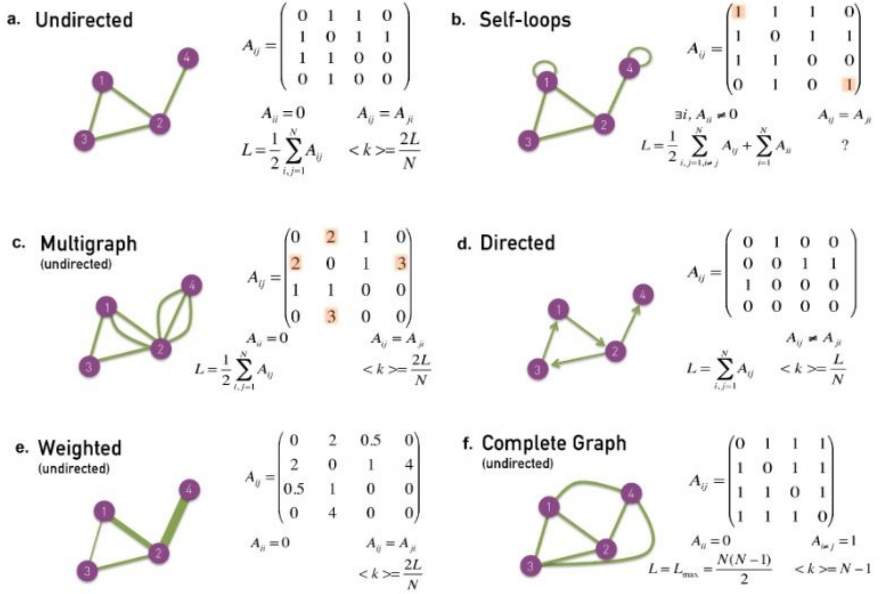
- Size: (N,E) (sum rows, columns in adj matrix)
- Degree: (k), and in-degree or out-degree

**Edge Level**

- Weight

**Graph Level**

- Avg degree:  2 * E/N  (<k>)
- Avg path length
- Diameter (longest shortest path)
- Density (ratio of edges to possible edges, 2e / (N*(N-1)))



# Statistics & Algos

**General Language**

- Graph theory vs Network Science (not hard and fast)
- Vertices/nodes (N), edges (E)
- Path (L), hops, cycles
- Undirected, directed, DAG, Bipartite
- Connectivity, cuts, component
- Hypergraphs

**Node Level**

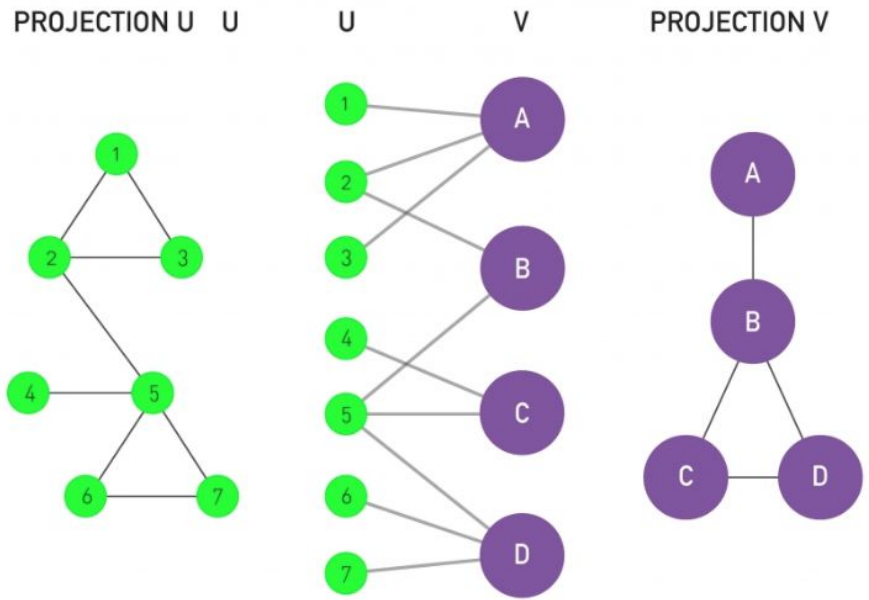
- Size: (N,E) (sum rows, columns in adj matrix)
- Degree: (k), and in-degree or out-degree

**Edge Level**

- Weight

**Graph Level**

- Avg degree:  2 * E/N  (<k>)
- Avg path length
- Diameter (longest shortest path)
- Density (ratio of edges to possible edges, 2e / (N*(N-1)))



# Statistics & Algos

**Clustering measures**

- e.g. for a node (2e / (k*(k-1))) where k is number of neighbors, and e is number of connections between neighbors
- e.g. 3 * number of triangles / number of connected triplets (Wasserman)

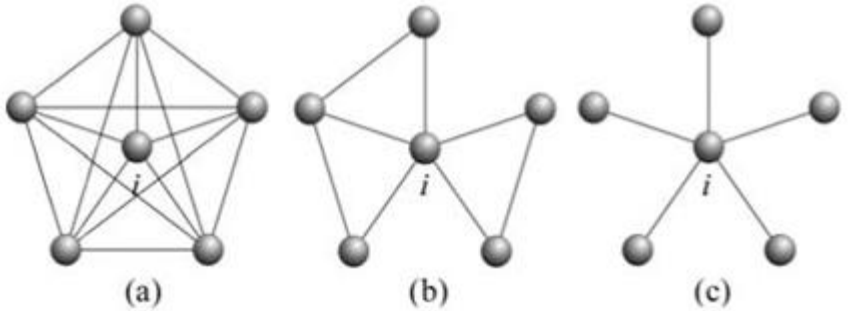



**Figure 4** - Example of three networks and respective clustering coefficients (see Eq. (1)). In (a), $cc_i = \frac{10(2)}{5(4)} = 1$ (the vertices around $i$ are fully connected), (b) $cc_i = \frac{3(2)}{5(4)} = 0.3$ and (c) $cc_i = \frac{0(2)}{5(4)} = 0$. The maximum number of edges among the neighbors of $i$ is given by $k_i(k_i - 1)/2$.

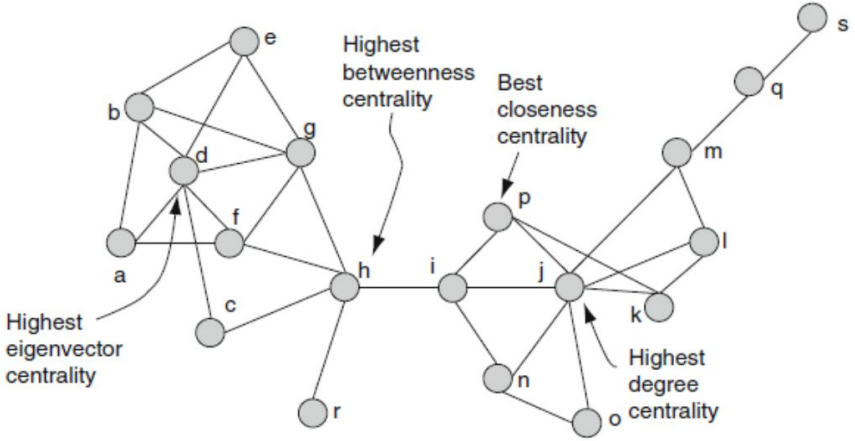
# Statistics & Algos

**Centrality measures**

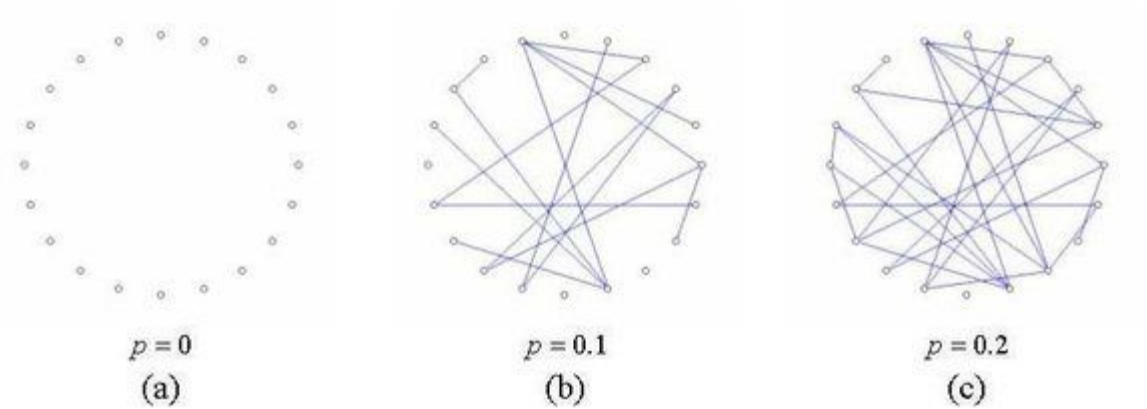
- http://en.wikipedia.org/wiki/Centrality
- Degree Centrality
- Closeness Centrality (1 / total distance)
- Betweenness Centrality (how many shortest paths pass thru…internet packet sending)
- Pagerank / HITS / Eigenvector
- Note: Can be used to flatten network / rank nodes (Discuss collecting data)

**Pagerank**

- http://www.cs.princeton.edu/~chazelle/courses/BIB/pagerank.htm
- PR(A) = (1-d) + d (PR(T1)/C(T1) + … + d(PR(Tn)/C(Tn)))
- C(X) is the vote
- PR(x) is the importance of the vote
- d is dampening
- Why hard to do in Hadoop?
- Clustering (i.e. community detection)
- Rank edges by betweenness and start removing edges to nodes in order



# Growth Modeling - Erdos Random Graph



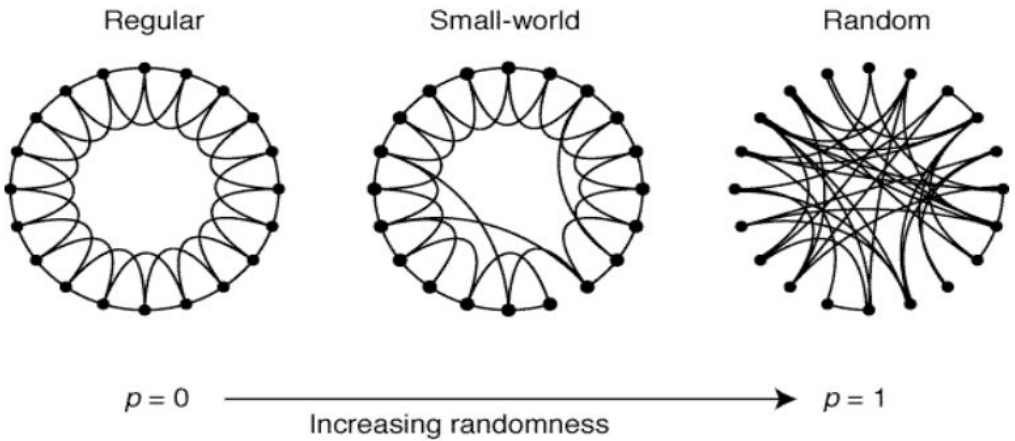

**Erdos**

- Review algorithm
- Connect a set of nodes with uniform probability (p)

**Descriptive Statistics**

- Properties depend on (Np) around value <k>
- Degree distribution is Poisson, giant component for large (N)
- Does not create hubs, triangle closures, Clustering coefficient approaches 0

# Growth Modeling - Watts-Strogatz



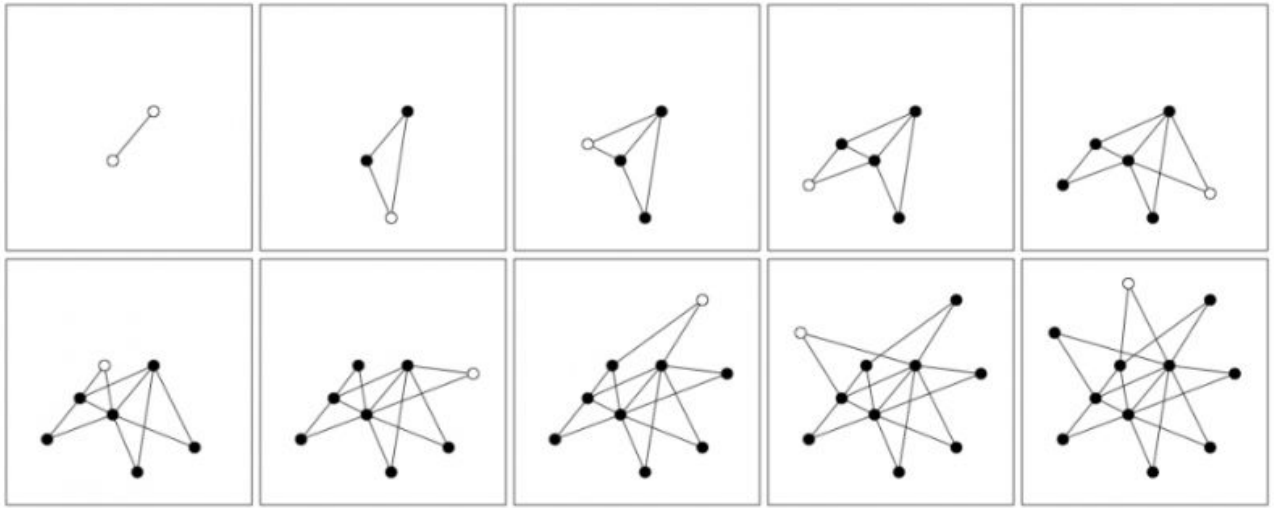

**Watts-Strogatz**

- Algorithm
- Every node has same number of edges in a ring lattice
- Random rewiring with fixed probability

**Descriptive Statistics**

- Creates locally clustered graph
- “Small World” is when the L between two random nodes is proportional to log(N)

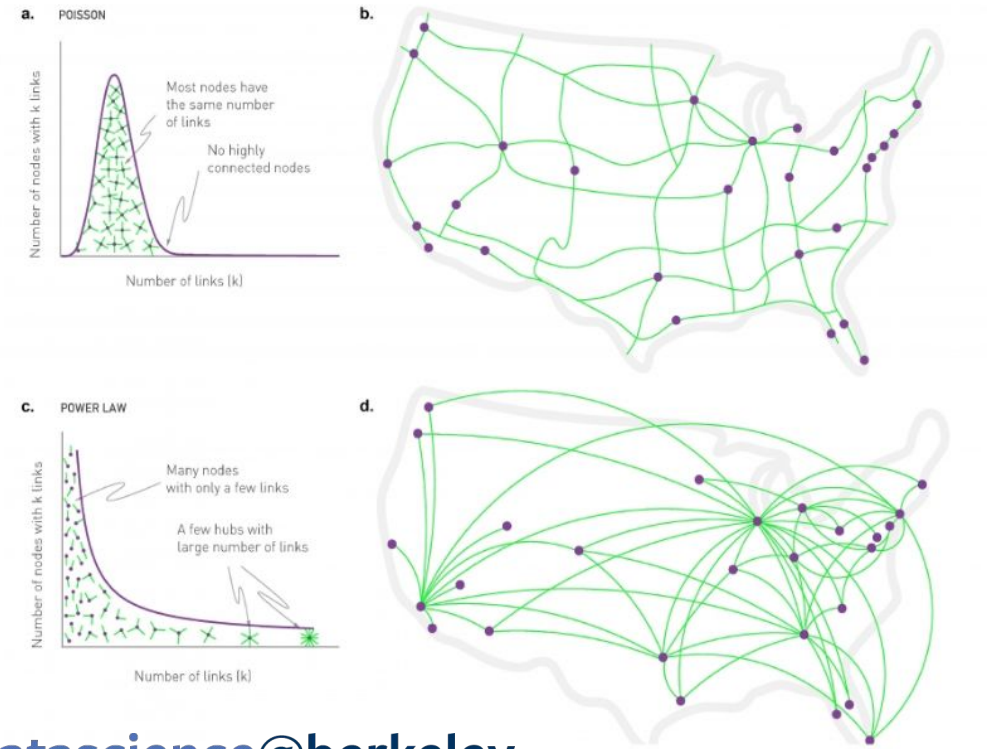
# Growth Modeling - Barabasi

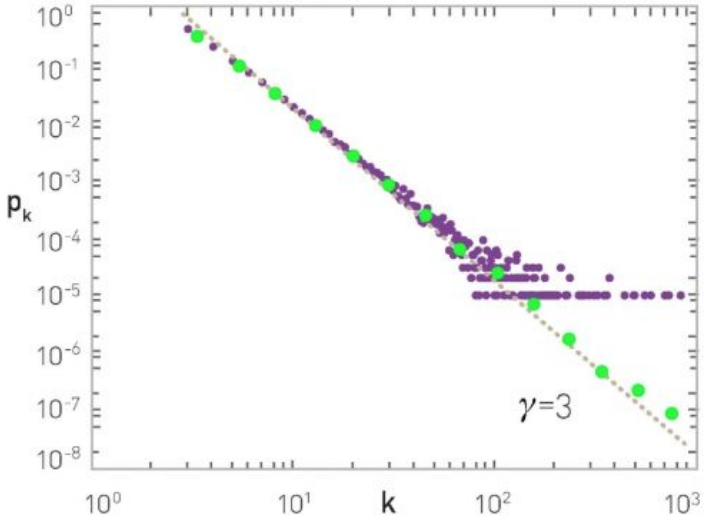


**Barabasi Preferential Attachment**

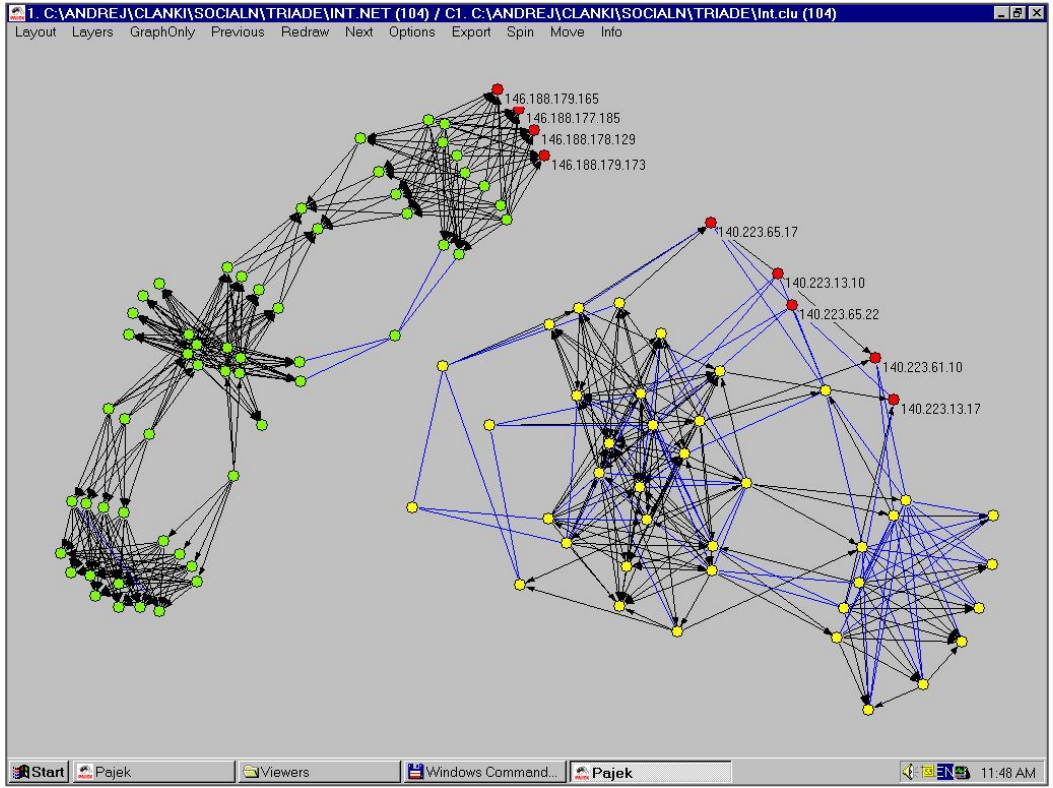
- p(k) is probability that a node has degree k
- log scale / power law (p(k) ~ k^-3)

# Growth Modeling - Barabasi



# Pajek


1. C:\ANDREJ\CLANKI\SOCIALN\TRIADE\INT.NET (104) / C1. C:\ANDREJ\CLANKI\SOCIALN\TRIADE\Int.clu (104)
Layout Layers GraphOnly Previous Redraw Next Options Export Spin Move Info
146.188.179.165
146.188.177.185
146.188.178.129
146.188.179.173
140.223.65.17
140.223.13.10
140.223.65.22
140.223.61.10
140.223.13.17
Start
Pajek
Viewers
Windows Command...
Pajek
11:48 AM

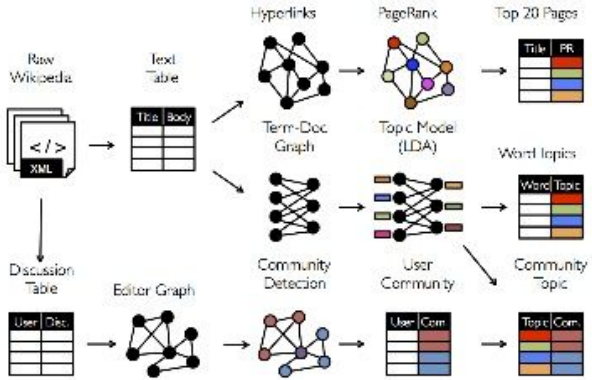
# GraphX

Scala

## GraphX (Apache Spark)

Offers a Graph API on top of Spark.

Enabling cross-world manipulations




ebiznext

# NetworkX

**NetworkX**

Stable (notes)

2.1 — January 2018
download | doc | pdf

Latest (notes)

2.2 development
github | doc | pdf

Archive

Contact

Mailing list
Issue tracker

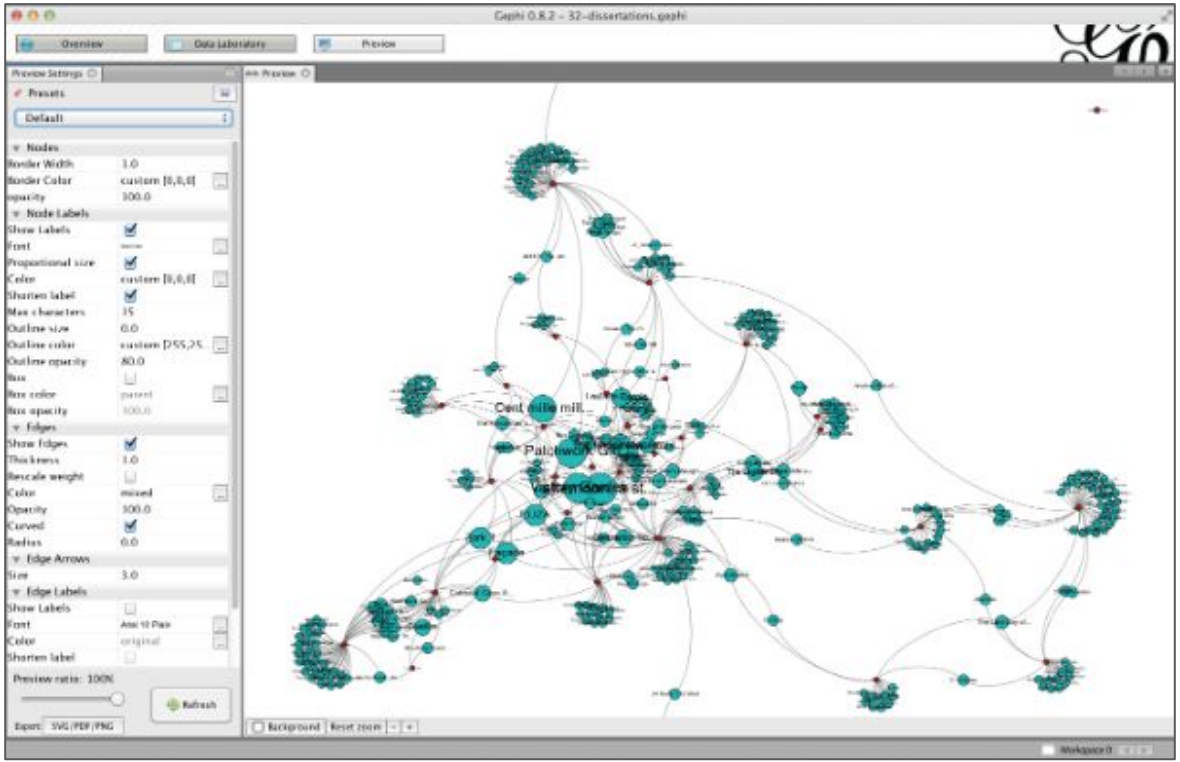
## Software for complex networks

NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.

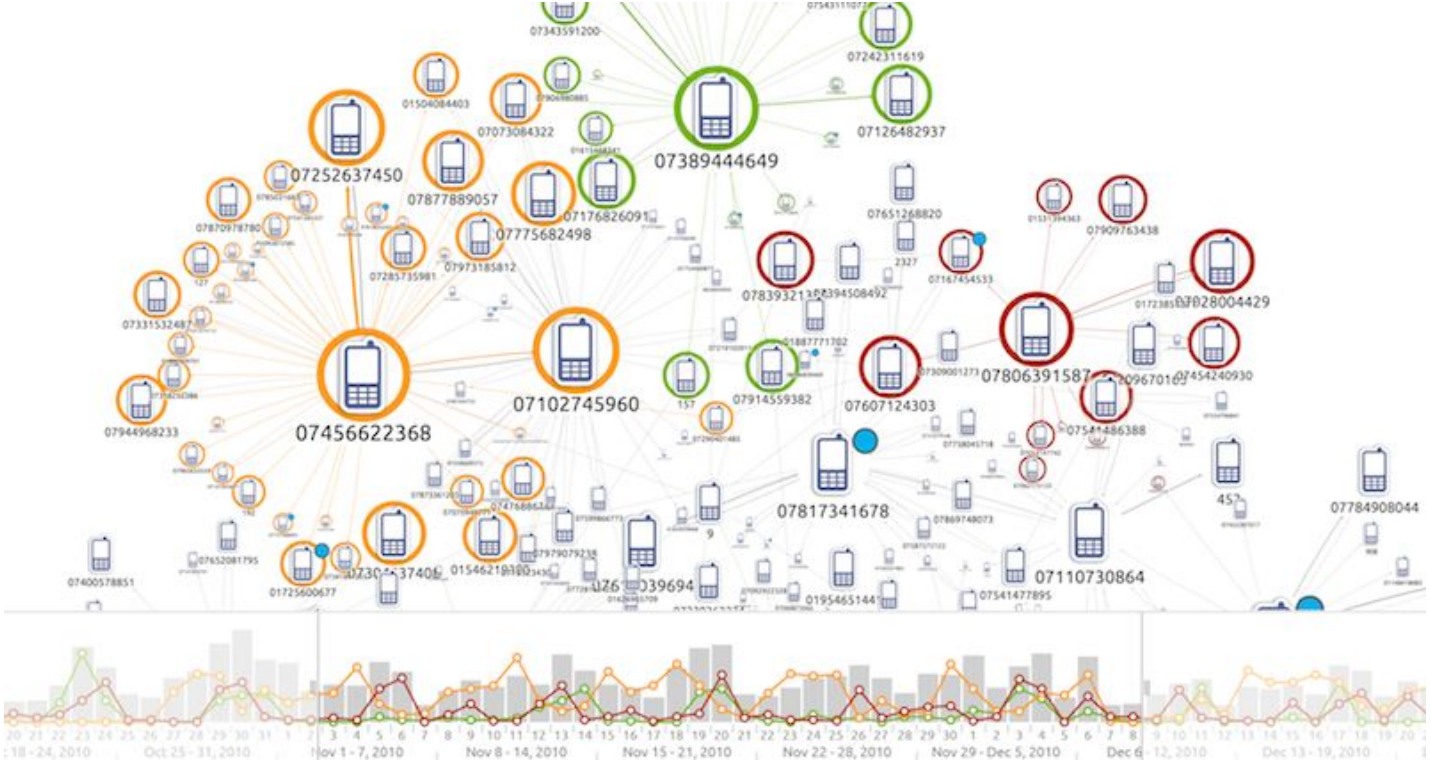
### Features

- Data structures for graphs, digraphs, and multigraphs
- Many standard graph algorithms
- Network structure and analysis measures
- Generators for classic graphs, random graphs, and synthetic networks
- Nodes can be "anything" (e.g., text, images, XML records)
- Edges can hold arbitrary data (e.g., weights, time-series)
- Open source 3-clause BSD license
- Well tested with over 90% code coverage
- Additional benefits from Python include fast prototyping, easy to teach, and multi-platform

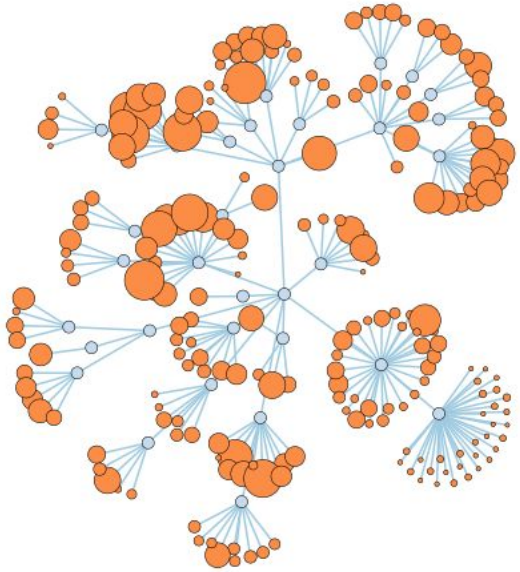
# Gephi


Gephi 0.8.2 - 32-dissertations.gephi
Overview
Data Laboratory
Preview
Preview Settings
Presets
Default
Nodes
Node Labels
Edges
Edge Arrows
Edge Labels
Refresh
Background
Reset zoom

# D3


07389444649
07252637450
07877889057
07176826091
07775682498
07102745960
07456622368
07944968233
07914559382
07607124303
07806391587
07817341678
07110730864
07126482937
07242311619
07909763438
07028004429
07784908044
Oct 25 - 31, 2010
Nov 1 - 7, 2010
Nov 8 - 14, 2010
Nov 15 - 21, 2010
Nov 22 - 28, 2010
Nov 29 - Dec 5, 2010
Dec 6 - 12, 2010
Dec 13 - 19, 2010

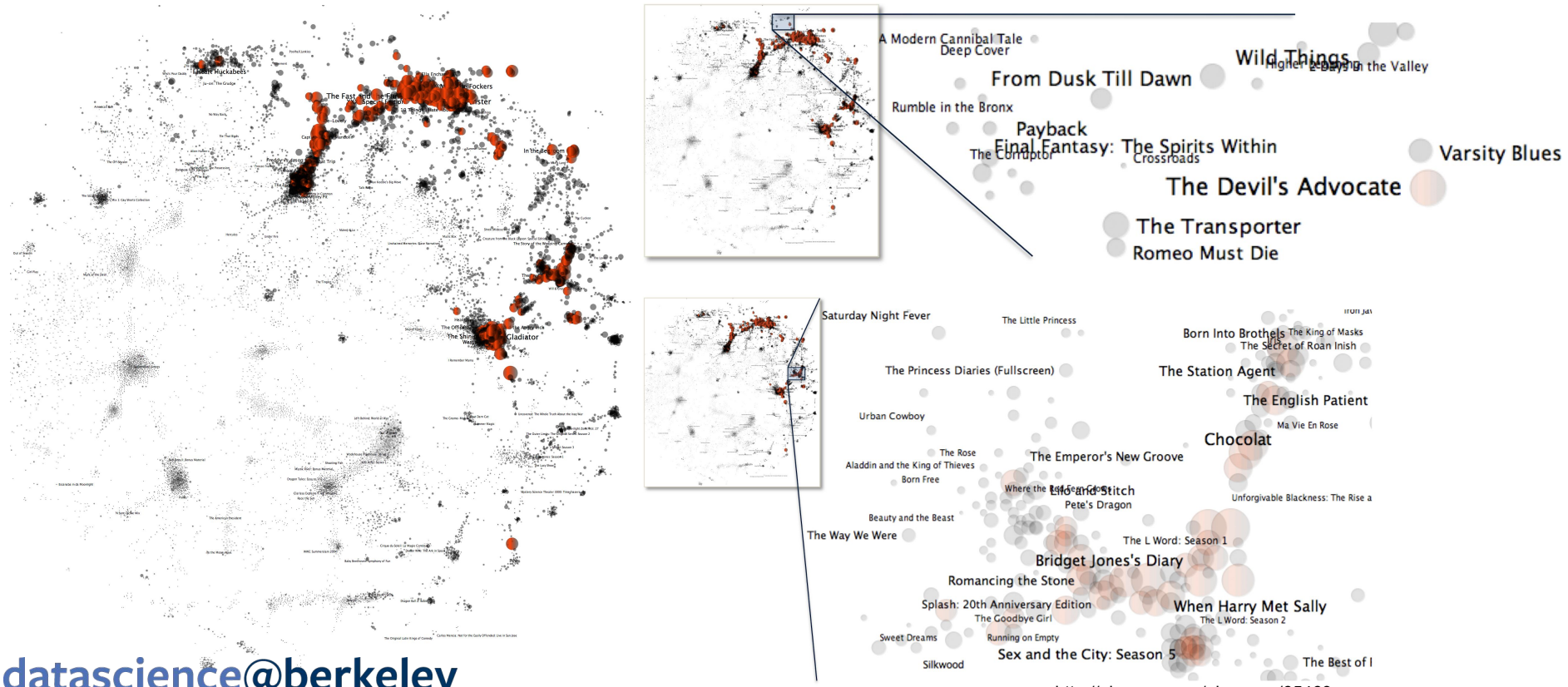
# D3 - “Force Directed”





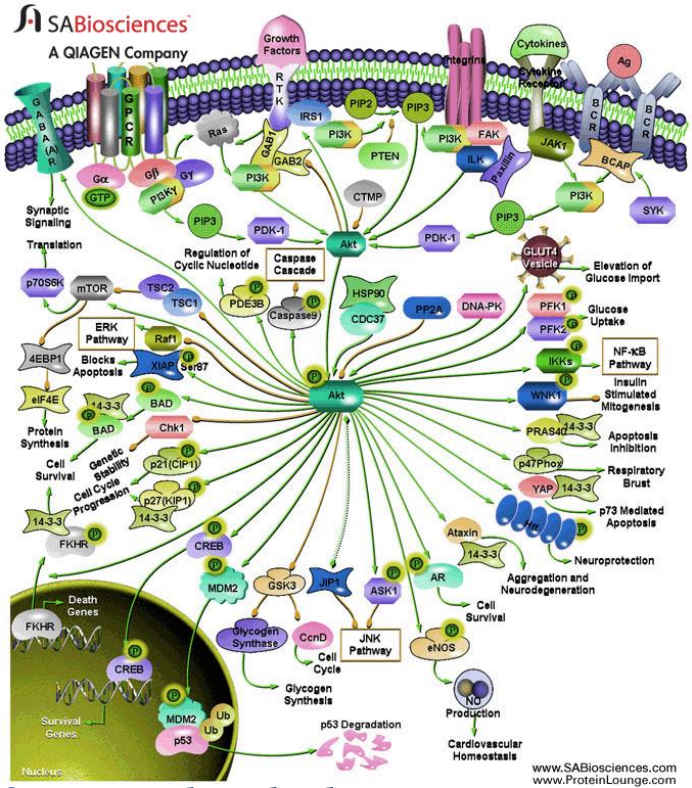
http://mbostock.github.io/d3/talk/20111116/force-collapsible.html

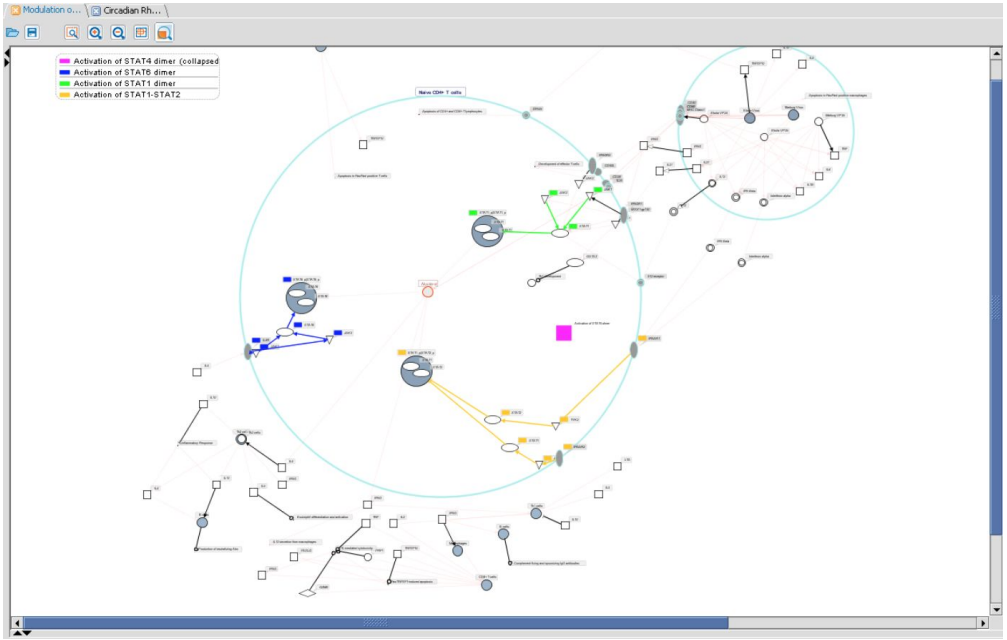
# Large Scale Visualization

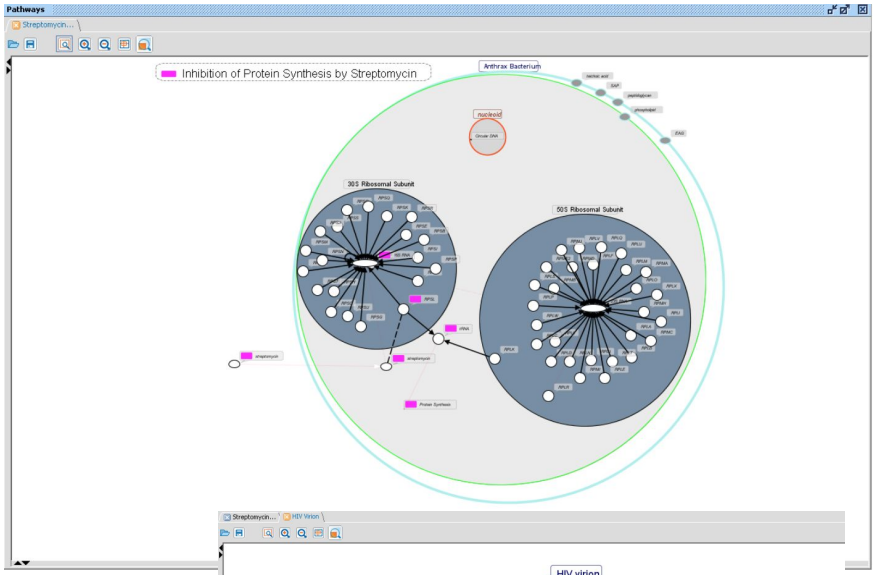


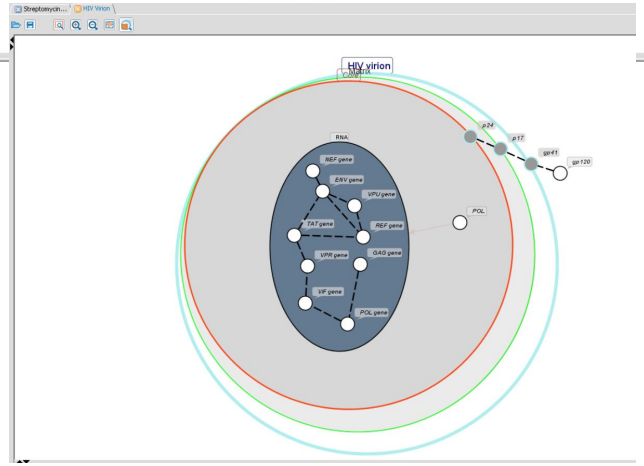

http://gigapan.com/gigapans/65469

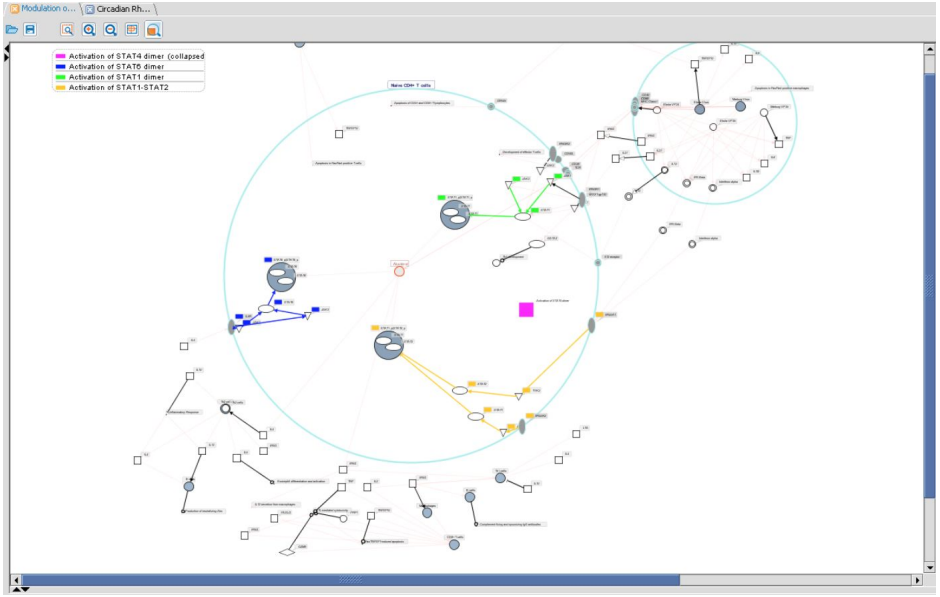
# Drug Discovery



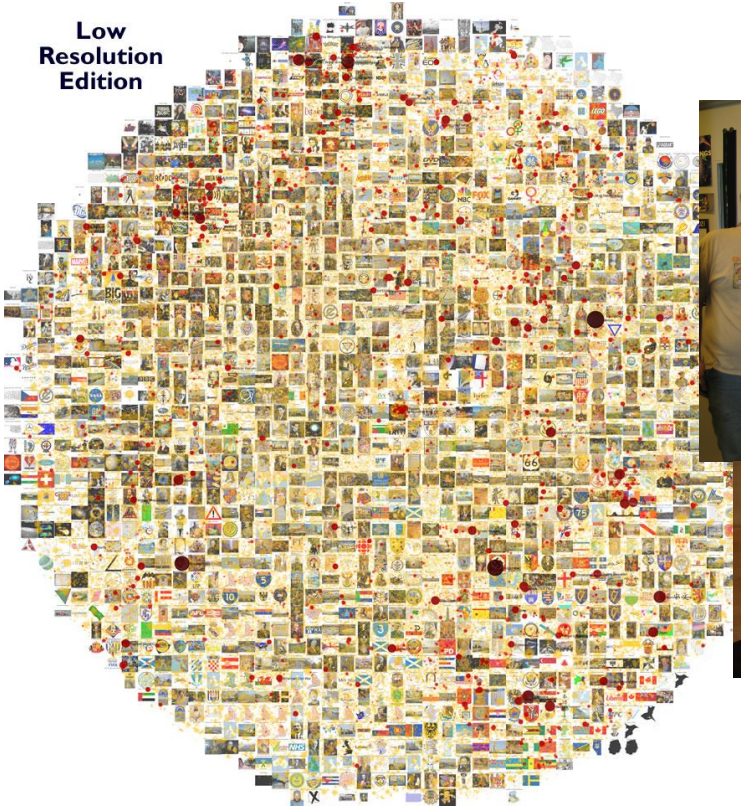




Pathways
Streptomycin...
Inhibition of Protein Synthesis by Streptomycin
Anthrax Bacterium
nucleoid
30S Ribosomal Subunit
50S Ribosomal Subunit
mRNA
streptomycin
Protein Synthesis

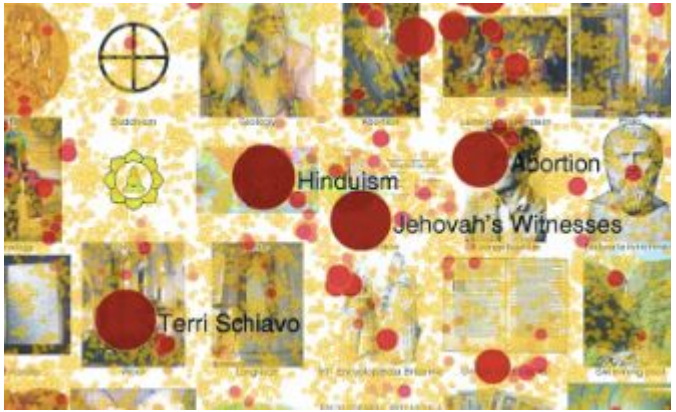

Streptomycin...
HIV Virion
HIV virion
Matrix
RNA
NEF gene
ENV gene
VPU gene
TAT gene
REF gene
VPR gene
GAG gene
VIF gene
POL gene
p24
p17
gp41
gp120
POL


Modulation o...
Circadian Rh...
Activation of STAT4 dimer (collapsed
Activation of STAT5 dimer
Activation of STAT1 dimer
Activation of STAT1-STAT2
Naive CD4+ T cells

# Large Scale Visualization



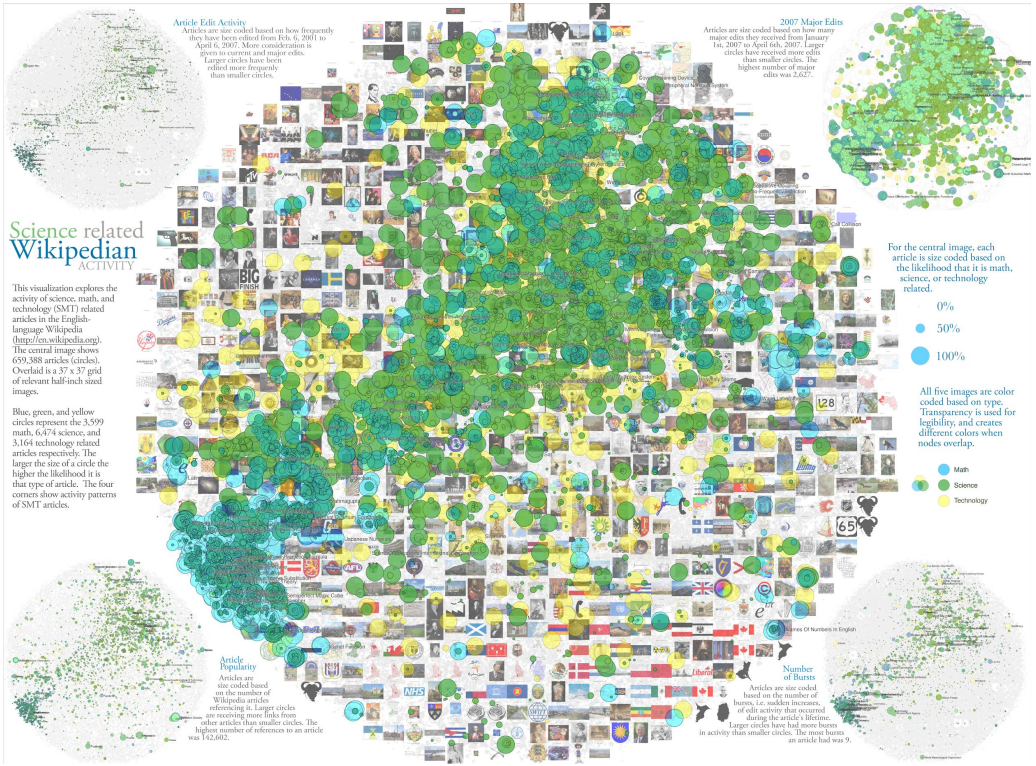


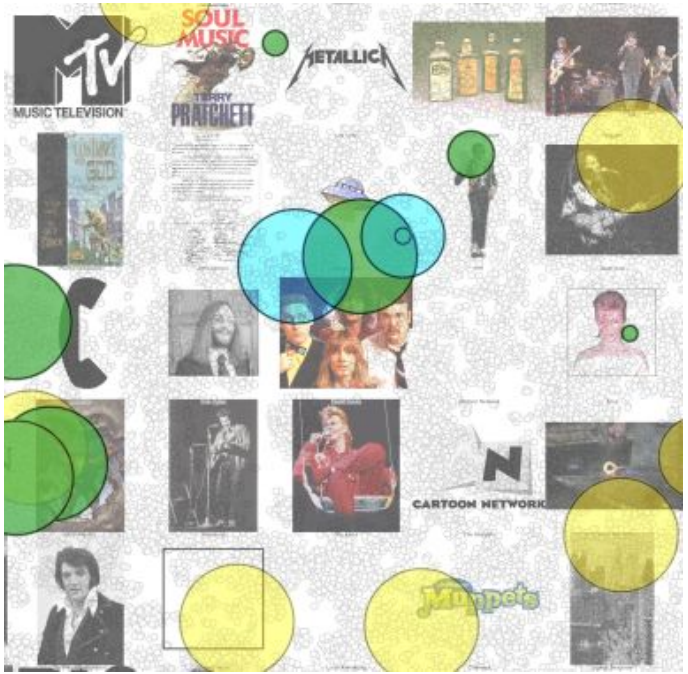




datascience@berkeley

http://gigapan.com/gigapans/4277

# Large Scale Visualization

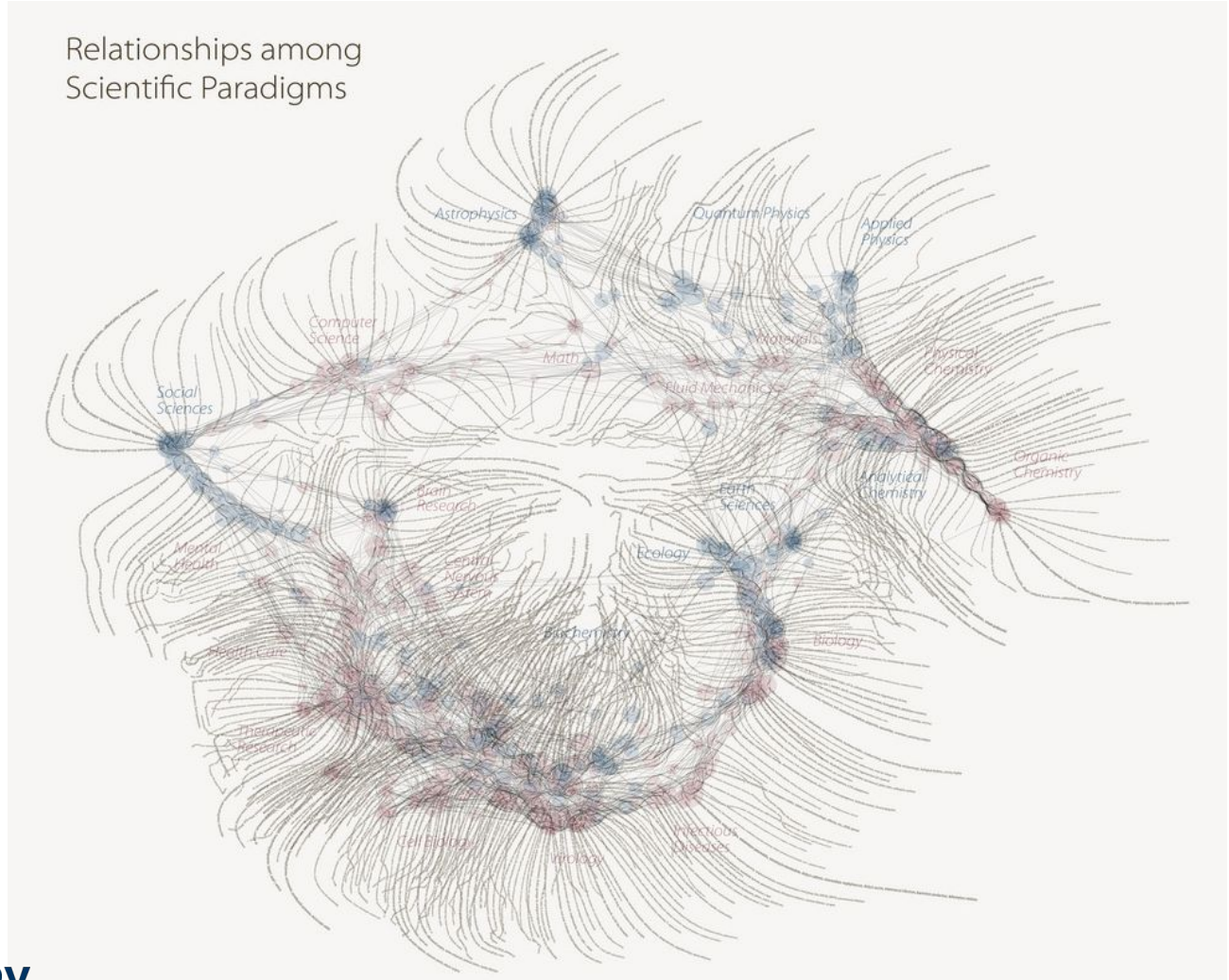





http://gigapan.com/gigapans/4305

# Large Scale Visualization




http://wbpaley.com/brad/mapOfScience/scienceMapFullColorPrint2_lowRes_b.jpg

# Large Scale Visualization