Complex Networks

Chen SiYu

Zhejiang University

sychen@zju.edu.cn

April 17, 2017

Why Complex Networks?

Concepts

- Community evolution
- Overlapping communities
- Directed networks
- Community characterization interpretation
- Modularities

Why Complex Networks?

- Unifying principles
- Dynamics

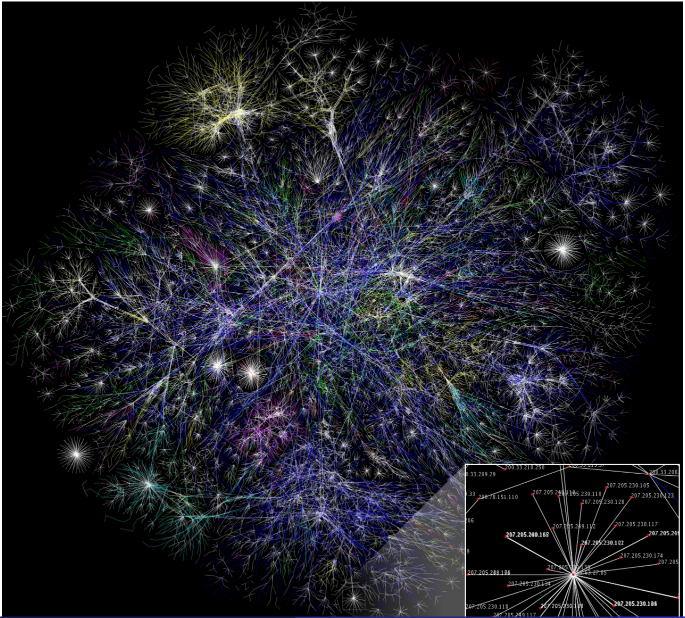
What for?

- prevailing products
- key users @Tc
- hot topics, ads
- followings, sortings
- clustering genes.

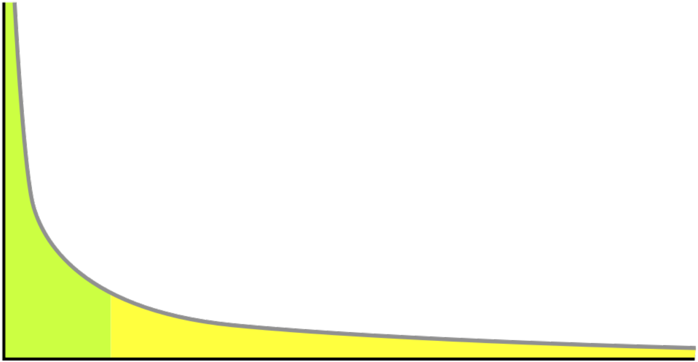
Definition by QIAN XueSen

- self-organizing
- sefl-similar
- attractor (simple,singular)
- small-world
- scale-free (power-law)

self similar



power law (long tail)

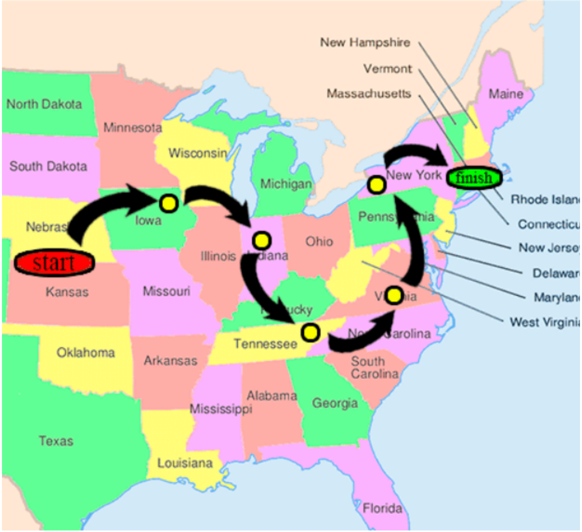


X frequency of words

Y priority when using

(Degree,Closeness∞,Harmonic Centrality)

small world experiment by Milgram, 1967

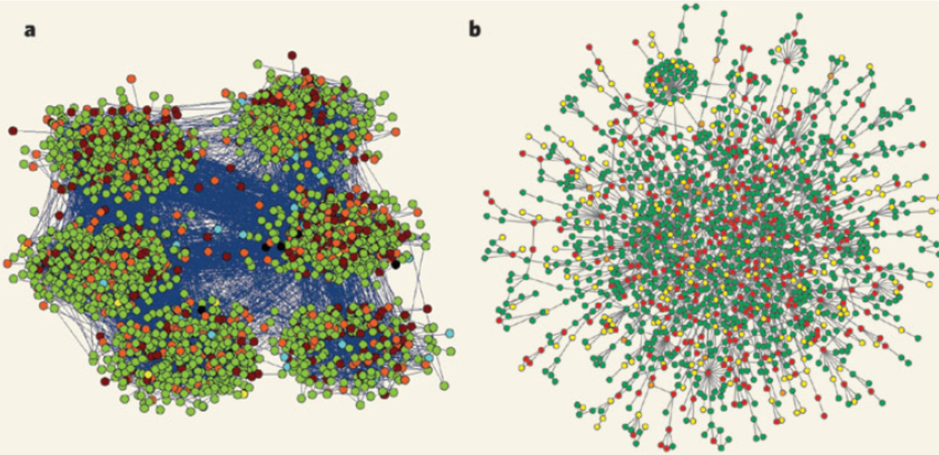


- 232/296
- 64
- 5.5-6

Characteristic

- Assortative Mixing: $E(i,j), \Delta(node_i, node_j)$ is small.
- Density $\frac{\text{avg deg}}{\text{complete deg}}$ smells cluster.
- Pearson corl-coeff: $cov(x,y) = E[(x - E[x])(y - E[y])], \frac{cov(x,y)}{\sigma_x \sigma_y}$
- Scalar attributes tends to be correlated.
- Probablistic properties: hidden vars behined time span.(Gaming)

Characteristic



a: school communities;

b: proteins in brewer's yeast.

- Clique, complete subgraphs.
- Density $\dfrac{\text{avg deg}}{\text{complete deg}}$
- Modularity: the extent of module-divided formation of networks. Defined as: measurement of extra edges aside of random connection.

Algorithms

- Heirachical
- Information theory
- Graph theory
- Other stupid fantacies

Erdos-Renyi probability based graph model.1959

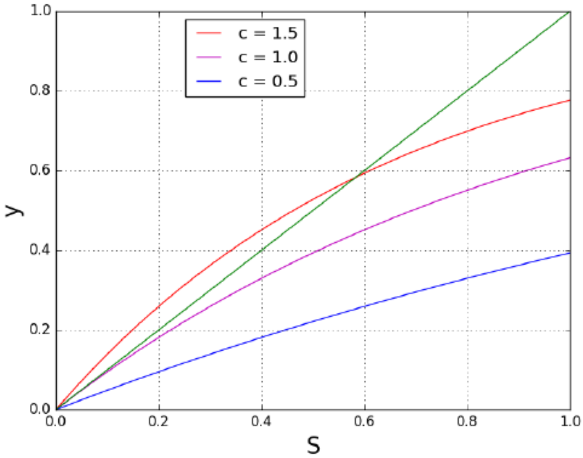
- $Pr(m) = \binom{\binom{n}{2}}{m} p^m (1-p)^{\binom{n}{2}-m}$
- $\mathbb{E}[m] = \binom{n}{2} p, \mathbb{E}[k] = \frac{2}{n}\binom{n}{2} p$
- $Pr(k) = \binom{n-1}{k} p^k (1-p)^{(n-1-k)}$

$Pr(k) \simeq \frac{c^k}{k!} e^c$ (Poisson)

Phase Transition

- $Pr(\neg e_{i,j}) = 1 - p, Pr(e_{i,j} \wedge \neg e_{j,compo}) = pu$
- $u = (1 - p + pu)^{n-1} = [1 - \frac{c(1-u)}{n-1}]^{n-1}$
- $\lim_{n\to\infty} u = e^{-c(1-u)}, Pr(e_{i,comp}) = 1 - u = S$
- $\lim_{n\to\infty} S = 1 - e^{-cS}$

Erdos-Renyi probability based graph model.



$$S = 1 - e^{-cS}, \text{if}, S > \frac{\ln(c)}{c}, c \geq 1$$

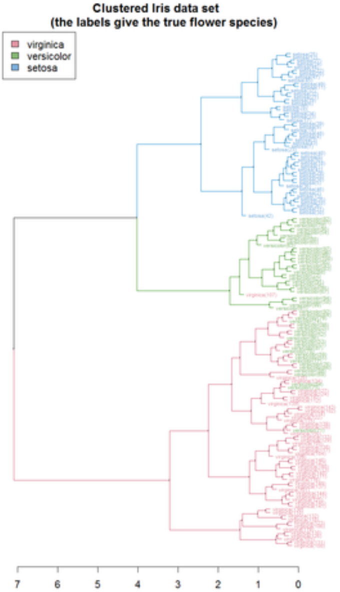
Girvan - Newman algorithm 2004
Sometimes wrongly called NG.
Some preceding concepts:

- Vertex bewteeneess
- Edge betweenness
- Pareto principle: 80/20 rule

Fundamental ideas:

- Power-law/Pareto principle yield isolation
- The minorities act as gate-nodes
- Removing those path reveals the communities.

Girvan - Newman algorithm



Dendrogram Breadth first

Latent Semantic Analysis 1988

$$D = \{d_1, d_2, .., d_N\} \tag{1}$$

$$W = \{w_1, w_2, .., w_M\} \tag{2}$$

$$Z = \{z_1, z_2, .., z_K\} \tag{3}$$

$$A_{coappearance} \text{ is } N \times M \tag{4}$$

$$A = U_{N\times r}\Sigma_{r\times r}V_{r\times M}^T \text{ (SVD/NMF)} \tag{5}$$

Probablistic Latent Semantic Analysis 2000, by Thomas Hofmann.

$$P(d_i, w_j) = P(d_i)P(w_j|d_i), \tag{6}$$

$$P(w_j|d_i) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \tag{7}$$

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \log P(d_i, w_i) \tag{8}$$

$$P(z_k|d_i, w_j) = \frac{P(z_k)P(d_i|z_k)P(w_j|z_k)}{P(d_i, w_j)} \tag{9}$$

$$\mathrm{E}[\mathcal{L}] = \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j) \sum_{k=1}^{K} \cdots \tag{10}$$

$$P(z_k|d_i, w_j) \log P(w_j|z_k)P(z_k|d_i) \tag{11}$$

Expectation Maximization

$$\sum_j P(w_j|z_k) = 1 \qquad (12)$$

$$\sum_i P(z_k|d_i) = 1 \qquad (13)$$

$$KKT : \frac{\partial L}{\partial P(w_j|z_k)} = 0, \frac{\partial L}{\partial P(z_k|d_i)} = 0 \qquad (14)$$

MASHA, HPLSA, Latent Dirichlet allocation

Dual problem

- $f(x), s.t. g(x) \leq 0$
- $L(x, \lambda) = f(x) + \lambda g(x), \lambda \geq 0$
- $argmax_{\lambda} L = f =$ initial problem
- we define original problem as : $argmin_{x} argmax_{\lambda} L(x, \lambda) s.t. \lambda \geq 0, g(x) \leq 0$
- $L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*)$
- Karush-Kuhn-Tucker conditions

Dual problem-KKT conditions

- $\lambda^* g(x^*) = 0$
- $\dfrac{\partial L}{\partial x} = 0$
- $\lambda \geq 0, g(x) \leq 0$

KKT is necessary and sufficient condition.

- $\max\limits_{\lambda} \min\limits_{x} L(x, \lambda), s.t. \lambda \geq 0, g(x) \leq 0$
- $\text{KKT} \rightarrow x = h(\lambda)$
- $\max\limits_{\lambda} L(h(\lambda), \lambda) s.t. \lambda \geq 0, S(\lambda) = 0$

Expectation Maximization

- $f(x;\theta) \to f(x, y; \theta)$
- randomize θ, y
- estimate $Q = q(y, \theta)$ (E-step)
- $\dfrac{\partial - L(Q)}{\partial\theta} = 0$,update θ (M-step)
- repeat until converge.

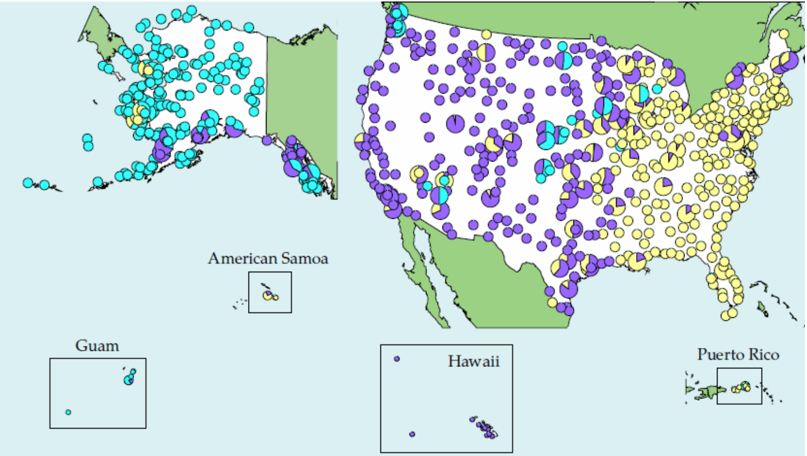
Expectation Maximization

- Jensen's inequality $f(\sum \lambda x) \leq \sum \lambda f(x)$
- $\sum \log \sum p \geq \sum \sum Q \log \frac{p}{Q} = l(\theta)$
- $l(\theta)$ is the lower bound.
- $f(E_z[\frac{p(x,z;\theta)}{Q(z)}]) \geq E_z[f(\frac{p(x,z;\theta)}{Q(z)})]$
- $\frac{p}{Q} = Const$ s.t. $\sum_z Q(z) = 1$
- $Q(z) = \frac{p(x,z)}{\sum p(x)} = p(z|x;\theta)$

Brian Ball, Brian Karrer, M. E. J. Newman 2011

- generative model
- use Z colors to annotate communities
- vertex v_i has Z params θ_{iz}
- $P(Edge(v_i, v_j)) = \sum_z \theta_{iz}\theta_{jz}$
- EM for θ

BKN is able to detect overlappings.



Latent community discovery network 2012 (Tsinghua). (PLSA-like)

- users with significant influence: core actors a_i.
- ordinary user u
- document of u, $d = a_1, ..., a_{|d|}$, network n
- count(a,d): occurrence of a in d.
- topics z

- how to get a list of core actors?
- $\sum_{z} P(z|d) = 1$,(mixing)
- $\sum_{a} P(a|z) = 1$ (topics are prevailing)
- $L = \sum_{d} \sum_{a} C(a, d) \log \sum_{z} P(a|z)P(z|d)$
- $R = \sum_{a1,a2} \sum_{z} \|p(z|a_1) - p(z|a_2)\|^2$ what?

- $Objective = \alpha(-L) + (1-\alpha)R$
- EM again.

AI DB DP GV NC

	Pairwise precision	Pairwise recall	Pairwise F1	Time cost(s)	Community size				
					C1	C2	C3	C4	C5
PLSA	0.276	0.238	0.265	19	243	199	244	256	299
k-means	0.257	**0.978**	0.406	569	2	19	1218	1	1
NG	Unavailable since this network is not fully interconnected								
BKN	0.156	0.306	0.206	799	852	100	80	126	83
LCDN	**0.456**	0.434	**0.445**	114	136	370	108	251	376

Table 1: Algorithm performance comparison on the DBLP co-authorship network

	Pairwise precision	Pairwise recall	Pairwise F1	Time cost(s)	Community size				
					C1	C2	C3	C4	C5
PLSA	0.309	0.245	0.273	18	221	252	242	213	294
k-means	0.258	0.979	0.409	432	1199	19	1	1	2
NG	0.253	**0.988**	0.403	8625	1217	1	1	1	2
BKN	0.287	0.480	0.359	755	764	102	104	125	127
LCDN	**0.501**	0.465	**0.483**	267	394	298	171	92	267

Table 2: Comparison on the fully interconnected DBLP co-authorship network

Latent community discovery network

1. Entertainment244 2. Leisure, 333
3. Finance, 297 4. Culture, 185
5. Media, 163

	Pairwise precision	Pairwise recall	Pairwise F1	Time cost(s)	Community size				
					C1	C2	C3	C4	C5
PLSA	0.627	0.567	0.596	1098	176	200	235	299	271
k-means	0.227	0.715	0.345	558	77	108	2	992	2
NG	0.201	0.873	0.326	85084	11	1124	21	10	15
BKN	0.528	0.478	0.502	4164	184	270	227	304	196
LCDN	**0.682**	0.611	**0.645**	2067	**282**	**185**	**221**	**277**	**216**

Table 3: Algorithm performance comparison on the WEIBO network

If you torture the data long enough, data will confess.

Page rank 1998.

- incoming-links: other $\rightarrow$ this site.
- Pagerank(site) $PR(site)$
- $PR(A) = \left(\frac{PR(B)}{L(B)} + \frac{PR(C)}{L(C)} + \cdots\right) d + \frac{1-d}{N}$
- $(1-d)$

- Idealrank 2009: local pagerank. far-away PR-values are simplified and unified.
- $\lim_{\substack{edge \to full \\ V_{local} \to V_{global}}} Idealrank = Pagerank$
- approxrank 2009: $approxrank \sim Idealrank$

Trust Network

- e-commerce
- resource sharing
- SQA,SDN
- propagenda
- promotion/ads

Zhang shaozhong. et al. 2012

- Interactions matrix A
- Successful Interactions matrix T
- Faliures F
- $Believe(v_i, v_j) = \sum \sum p(v_i, v_j) \log \frac{p(v_i, v_j)}{p(v_i)p(v_j)}$
- v: success or failure.

Mutual information

$$I(X;Y) = \sum_y \sum_x p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

if X, Y, iid. then $I = 0$

else $I > 0$

Trust Network

To find an optimal trusty path:

- Floyd
- Dijkstra

How to find communities?

How to find communities?

- interconnectivity in v $C_v = \dfrac{2t_v}{k_v(k_v-1)}$
- $\mathrm{E}[C_v] = C = \dfrac{\sum_{n=1}^{\infty} C_v}{n}$
- characteristic path length: $\dfrac{\sum_i \frac{\sum_j MinDist(i,j)}{n-1}}{n}$

How to find communities?

- $f = a\left(\sum Believe\right) + bC|_m$
- cut m weak interactions. (heirachical?)
- NP hard!
- a heuristic algorithm:

1 cut m edges $e_1, .., e_m$ that maximize f.

2 add e'_m that maximize f, if $e'_m = e_m$ over.

3 cut one edge that maximize f, then go to 2.

End

Thank you, gays.

- a
- b

Paragraphs of Text

Sed iaculis dapibus gravida. Morbi sed tortor erat, nec interdum arcu. Sed id lorem lectus. Quisque viverra augue id sem ornare non aliquam nibh tristique. Aenean in ligula nisl. Nulla sed tellus ipsum. Donec vestibulum ligula non lorem vulputate fermentum accumsan neque mollis.

Sed diam enim, sagittis nec condimentum sit amet, ullamcorper sit amet libero. Aliquam vel dui orci, a porta odio. Nullam id suscipit ipsum.

Bullet Points

- Lorem ipsum dolor sit amet, consectetur adipiscing elit
- Aliquam blandit faucibus nisi, sit amet dapibus enim tempus eu
- Nulla commodo, erat quis gravida posuere, elit lacus lobortis est, quis porttitor odio mauris at libero
- Nam cursus est eget velit posuere pellentesque
- Vestibulum faucibus velit a augue condimentum quis convallis nulla gravida

Blocks of Highlighted Text

Block 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer lectus nisl, ultricies in feugiat rutrum, porttitor sit amet augue. Aliquam ut tortor mauris. Sed volutpat ante purus, quis accumsan dolor.

Block 2

Pellentesque sed tellus purus. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Vestibulum quis magna at risus dictum tempor eu vitae velit.

Multiple Columns

Heading

1. Statement
2. Explanation
3. Example

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer lectus nisl, ultricies in feugiat rutrum, porttitor sit amet augue. Aliquam ut tortor mauris. Sed volutpat ante purus, quis accumsan dolor.

Table

Treatments	**Response 1**	**Response 2**
Treatment 1	0.0003262	0.562
Treatment 2	0.0015681	0.910
Treatment 3	0.0009271	0.296

Table: Table caption

Theorem

Theorem (Mass–energy equivalence)

$$E = mc^2$$

Verbatim

Example (Theorem Slide Code)

```
\begin{frame}
\frametitle{Theorem}
\begin{theorem}[Mass--energy equivalence
$E = mc^2$
\end{theorem}
\end{frame}
```

Figure

Uncomment the code on this slide to include your own image from the same directory as the template .TeX file.

Citation

An example of the `\cite` command to cite within the presentation:

This statement requires citation [Smith, 2012].

References

John Smith (2012)
Title of the publication
Journal Name 12(3), 45 – 678.

The End