What do features tell us about the underlying network?

Matteo Marsili (ICTP)
G. Bianconi (ICTP), S. Franz (LPTMS), P.Pin (Siena)

Data: network + features

- network is sometimes harder to observe than features
- inference: network ⇔ features?
- Choice vs opportunity bias in social networks
- An indicator for the relevance of features

Homophily:

"a contact between similar people occurs at a higher rate than among dissimilar people" (McPherson et alii, 2001)

Homophily is pervasive along many dimensions of diversity: race, age, sex, religion, profession. . . (Marsden, 1988)

Homophily influences behavior:

- formation and spread of opinions
- individual behavior (job search, investment, education)
- social behavior (voting, public goods)

Welfare implications

Definition of homophily

Coleman (1958) defines an index of inbreeding homophily of group i:

$$H_i \equiv \frac{q_i - p_i}{1 - p_i}$$

(F-statistics)

where

- p_i is the ratio of type i in the population
- q_i is the average ratio of i type in i's social ties

This measure considers a single dimension (race, education. . .)

Opportunity–based (OBH) and choice–based (CBH) homophily

What are the causes of homophily?

- Opportunity–based: it is due to opportunities
  - spatial segregation (race, census. . .)
  - different loci of activity (education, religion. . .)
  - difficulties in communication (language, culture. . .)
- Choice–based: it is due to individual choices
  - because of common interests and behavior
  - it is not necessarily the choice of one individual, but the effect of the aggregate choices

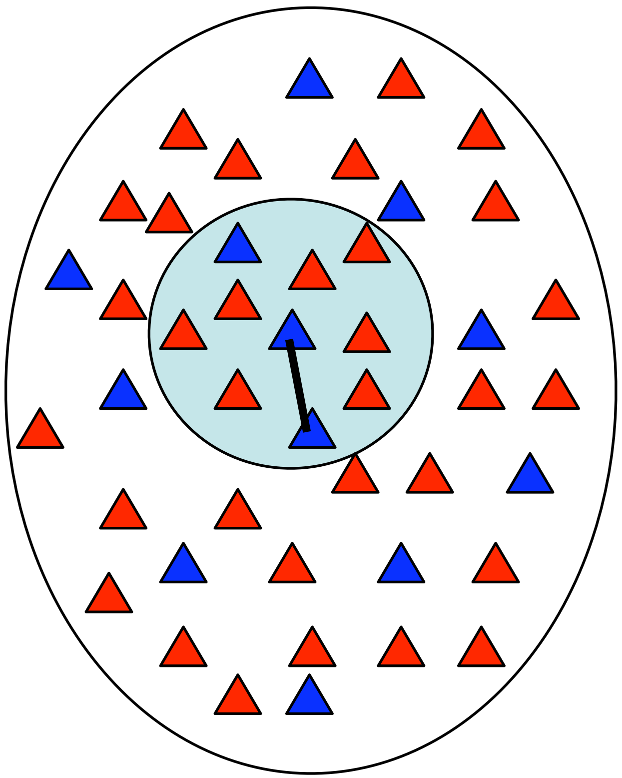
Literature

- Sociology
  - McPherson & Smith–Lovin (1987): distinguish OBH from CBH in friendships (data analysis on 457 questionnaires in Nebraska)
  - Moody (2001): Add Health schools – discusses the difference between OBH and CBH but is not able to disentangle
- Economics
  - Schelling (1971), Vinkovič and Kirman (2006): CBH influences OBH, across time
  - Bisin, Topa and Verdier (2004): disentangle dimensions of homophily
  - Currarini, Jackson and Pin (2007): strong non–linearity
- Physics
  - Jego and Roehner (2007): choice–based homophily is due to aggregate behavior

No quantitative method to distinguish OBH and CBH

The problem

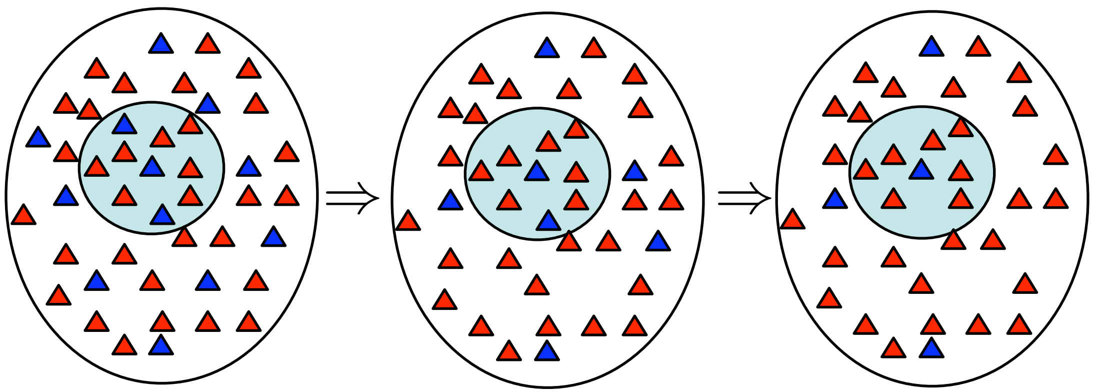
- Observable choices unobservable opportunities (neighborhood)
- infer underlying social network from choices
- not for single link: statistical tendency
- e.g. academic tracking in US schools



Intuition: Density dependence

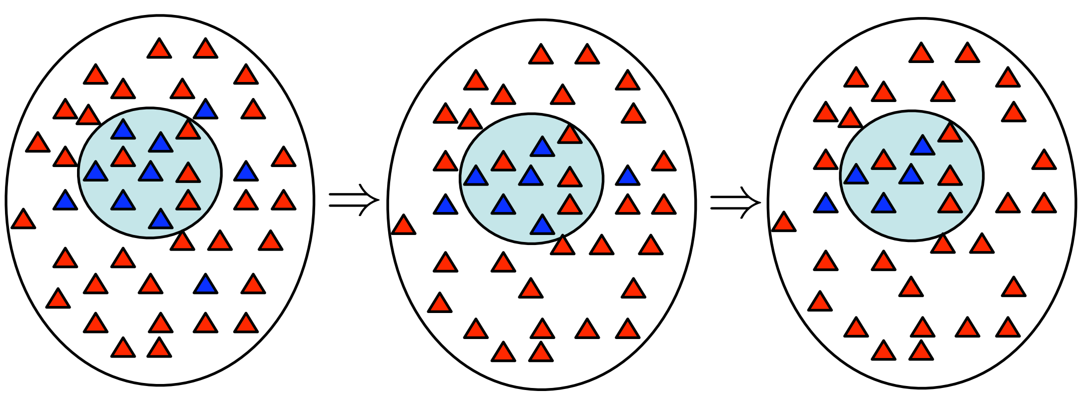
Large population with a fraction p of minority individuals
Finite neighborhood

Even with choice homophily, if there is no opportunity bias then



$$\lim_{p \to 0} H(p) = 0$$

If there is opportunity bias:



$$\lim_{p \to 0} H(p) > 0$$

p small: $H(p) \simeq A + Bp + O(p^2)$

A > 0 is an indicator of opportunity bias

- N individuals, pN of minority type
- Each i has a neighborhood of K others
- Neighborhood of minority i has a fraction

$$\bar{p} = \pi + (1 - \pi)p$$

of minority j's

- A minority j is chosen from the neighborhood x > 0 times more likely than a majority j to form a link
- Each individual i form k links with other j's

$$\pi \quad = \quad \text{measures opportunity bias} = \pi(A, B)$$

$$x \quad = \quad \text{measures choice bias} = x(A, B)$$

Data

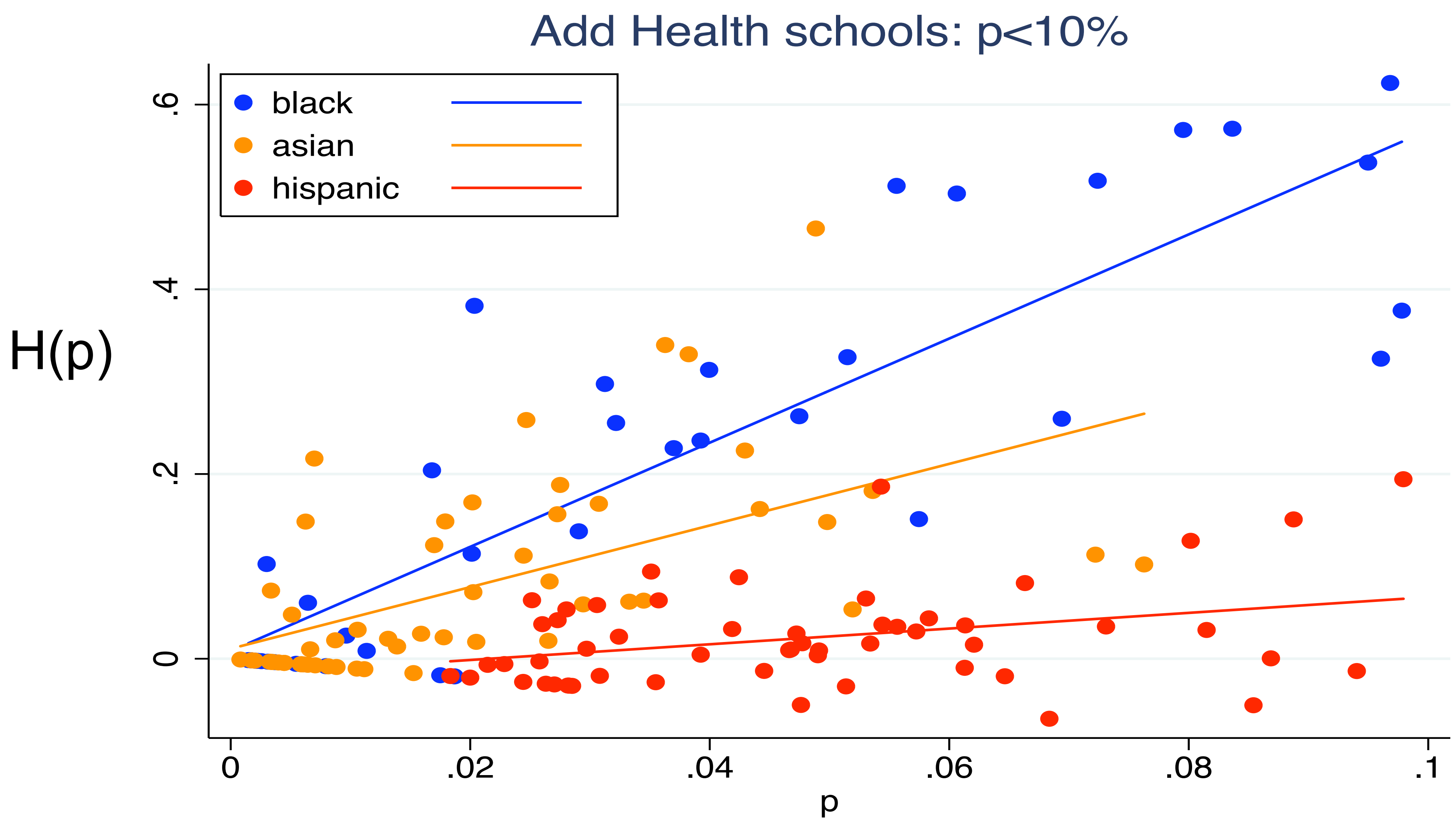
1- friendship in US schools

Add Health data: 1994 survey on 84 high-schools in US

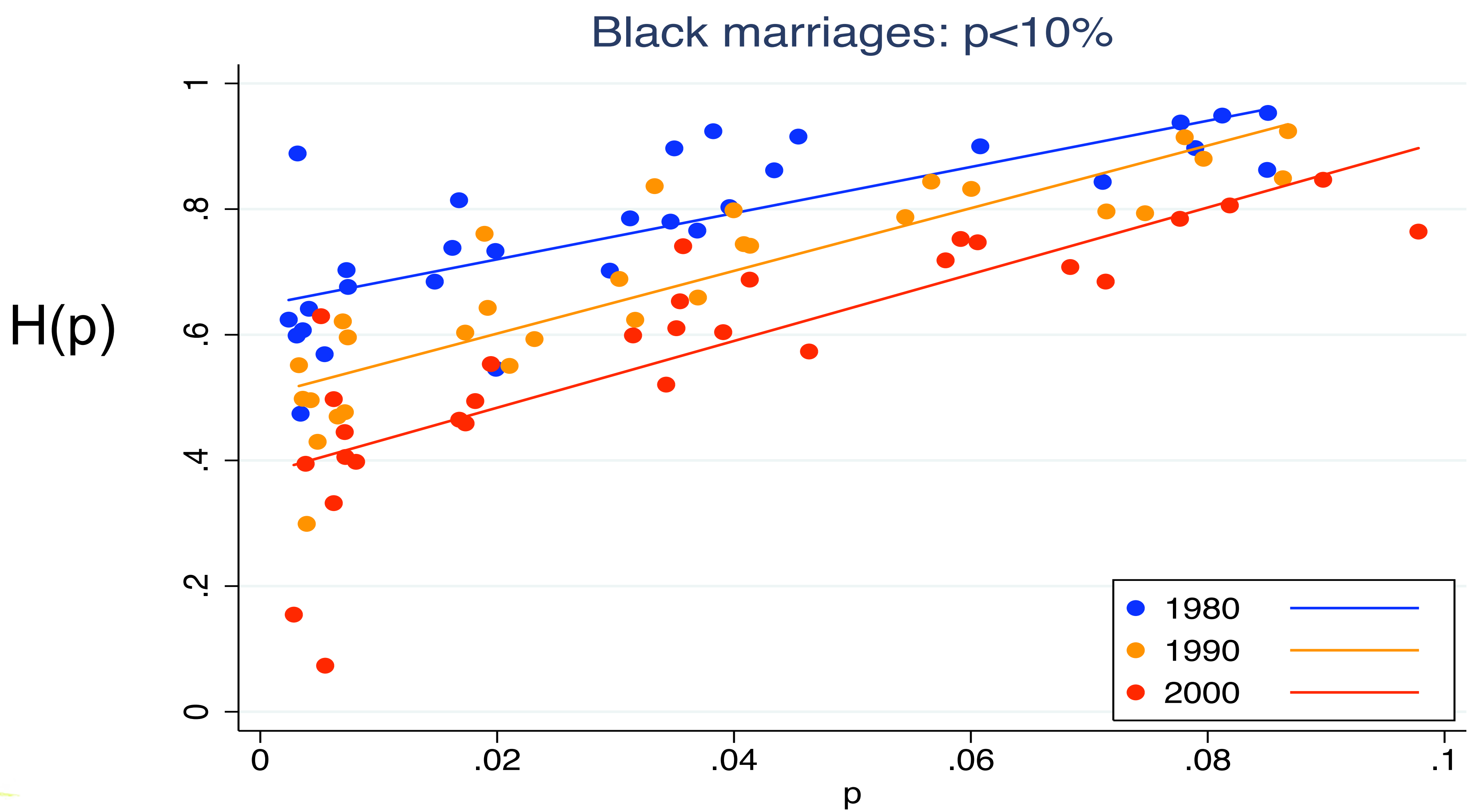
2- marriages in US

The Integrated Public Use Microdata Series (IPUMS) surveys on marriages in the 51 American States from years 1980, 1990 and 2000

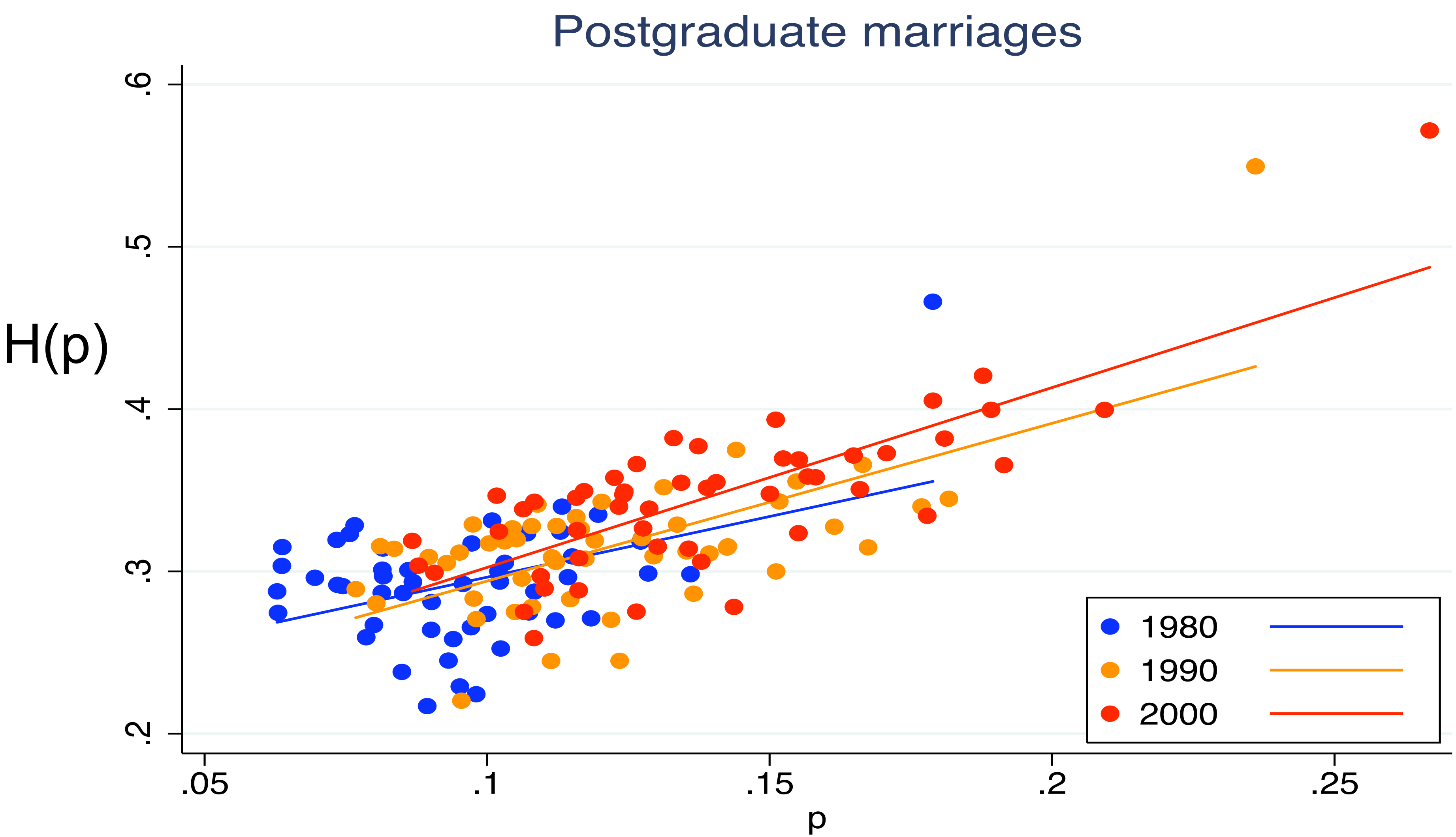
Add Health Data



IPUMS Data: black minority



IPUMS Data: education



Results: Table

IPUMS marriages in US states									
no threshold	n	B	±	A	±	x	±	π	±
Top Educ. 1980	51	0.7	0.4	0.22	0.04	1.2	0.7	0.113	0.04
Top Educ. 1990	51	1.0	0.3	0.20	0.04	1.5	0.5	0.089	0.02
Top Educ. 2000	51	1.1	0.3	0.19	0.04	1.7	0.4	0.081	0.02
10% threshold									
Black 1980	30	3.7	1.1	0.65	0.06	29.4	13.4	0.057	0.02
Black 1990	30	5.0	1.0	0.50	0.04	20.1	5.4	0.046	0.01
Black 2000	30	5.3	1.4	0.38	0.06	13.7	4.3	0.040	0.01
Asian 1980	50	8.3	4.0	0.38	0.06	21.9	11.3	0.026	0.01
Asian 1990	50	4.8	2.3	0.45	0.06	15.6	8.3	0.046	0.02
Asian 2000	49	5.5	2.2	0.47	0.06	19.1	8.7	0.042	0.02
Native 1980	50	6.8	1.6	0.16	0.04	9.6	2.5	0.018	0.01
Native 1990	50	4.5	1.3	0.14	0.04	6.2	1.8	0.023	0.01
Native 2000	49	4.1	1.3	0.17	0.04	5.9	2.0	0.029	0.01
Hispanic 1980	50	6.5	2.7	0.41	0.06	18.4	8.4	0.034	0.01
Add Health friendships in US schools									
Black	39	5.6	1.0	0.01	0.39	5.7	4.7	0.001	0.06
Asian	56	3.3	1.3	0.01	0.04	3.4	1.3	0.002	0.01
Hispanic	55	0.9	0.7	-0.02	0.04	0.8	0.6	-0.010	0.02

Results: Opportunity (OBH) and choice–based (CBH) homophily

1. In marriages, OBH is stronger for *top educated* people than for any racial minority, but CBH is much weaker
2. In marriages, OBH and CBH decrease for Blacks between 1980, 1990 and 2000 (no time–dependence for the other races and for *top educated* people)
3. School friendships do not exhibit OBH (compared to the school population), while marriages do
4. CBH is much stronger for marriages than for friendships
5. Both are strictly race–dependent
   - Blacks exhibit the strongest CBH and (in marriages) OBH
   - Hispanics exhibit the lowest values of both (~ 0 in schools)

Extensions:

- Opportunity and choice across other dimensions (religion, wealth, ...) and other countries

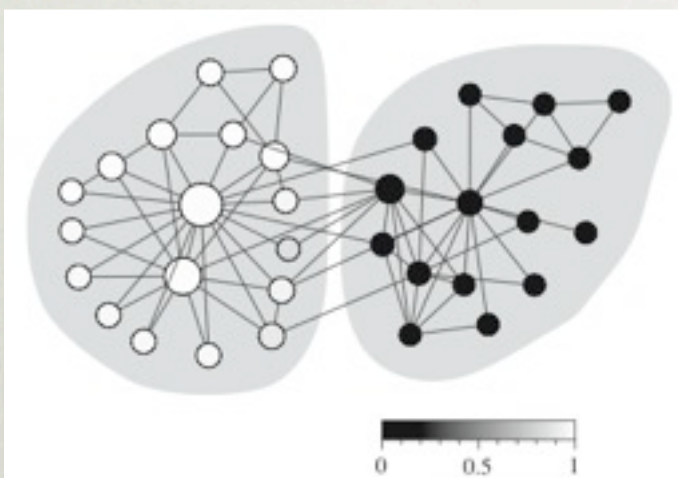
- Opportunity and choice in other contexts (scientific collaborations, R&D partnership, trade, ...)

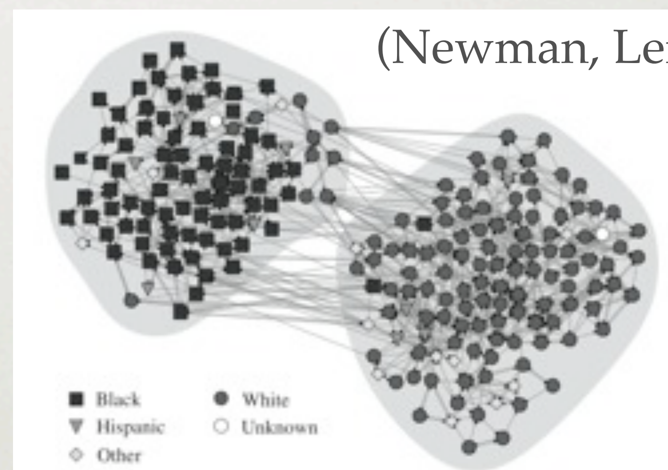
- How do choices bias opportunities over time? (e.g. what is the origin of OBH in dynamic models of social networks?)

How relevant is a given feature for a network?

Network -> features e.g. community detection

- Tens of algorithms (and authors)
- Performance:
  benchmarks + known classification





(Newman, Leicht PNAS 104, 2007)

- Algorithm dependent outcome

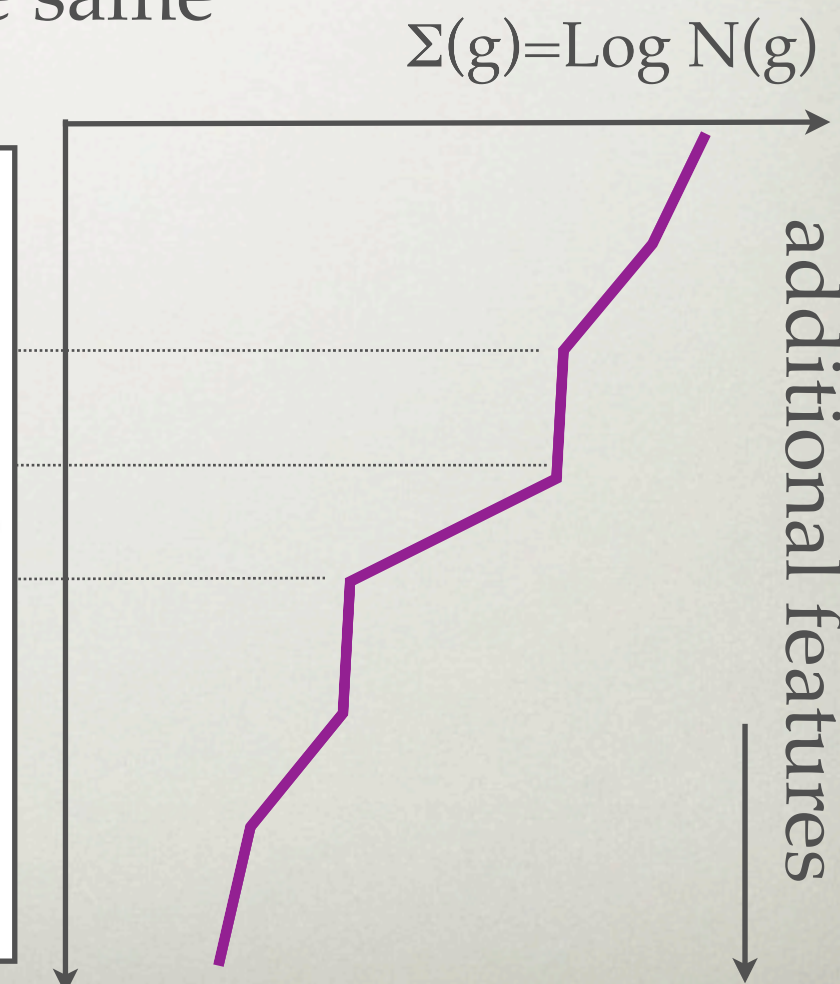
features -> Network e.g. known classification

- How much does an assignment of nodes into classes constrains the number of possible networks?
  (entropy of network ensembles)
- Universal answer
- information bound on feature detection algorithms
- reveal hidden statistical regularities

Entropy of network ensembles

(G. Bianconi '08)

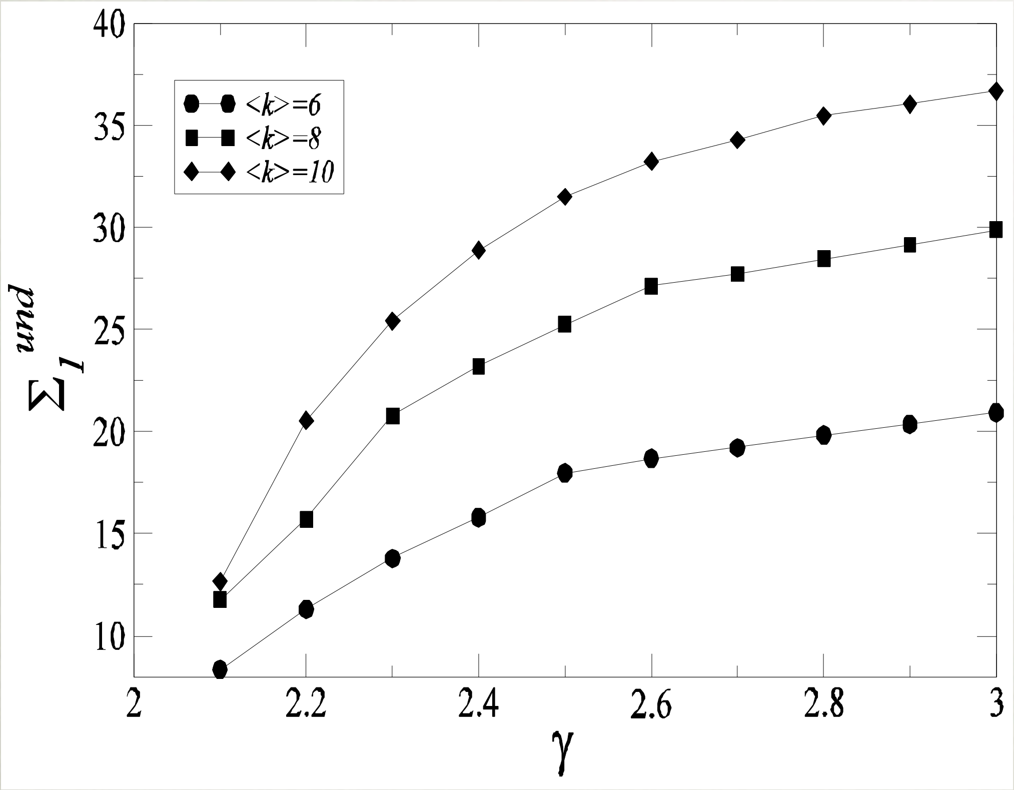
For a given network g with n nodes, how many networks are there with the same



The relevance of a feature is measured by how much its addition decreases the (log of the) number of networks in the ensemble, i.e. the entropy

For example: fixed degree sequence

n=4 $\Sigma(k_i=2,2,2,2)=1$, $\Sigma(k_i=1,1,2,2)=2$



Knowing the degree distribution:
e.g. scale free graphs

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

G. Bianconi, Europhys. Lett 2008)

The indicator

- Fixed degree sequence g + feature q

$$\Theta_{g,q} = \frac{\langle \Sigma_{\phi(g,\pi(q))} \rangle_\pi - \Sigma_{\phi(g,q)}}{\sqrt{\langle \delta \Sigma^2_{\phi(g,\pi(q))} \rangle_\pi}}$$

- π(q) random permutation of feature across nodes
- MC estimate of $\langle \ldots \rangle_\pi$ on M samples
  $\Rightarrow$ confidence interval at p=1/M

Feature = community

Σ(g,A) = log Number of networks with

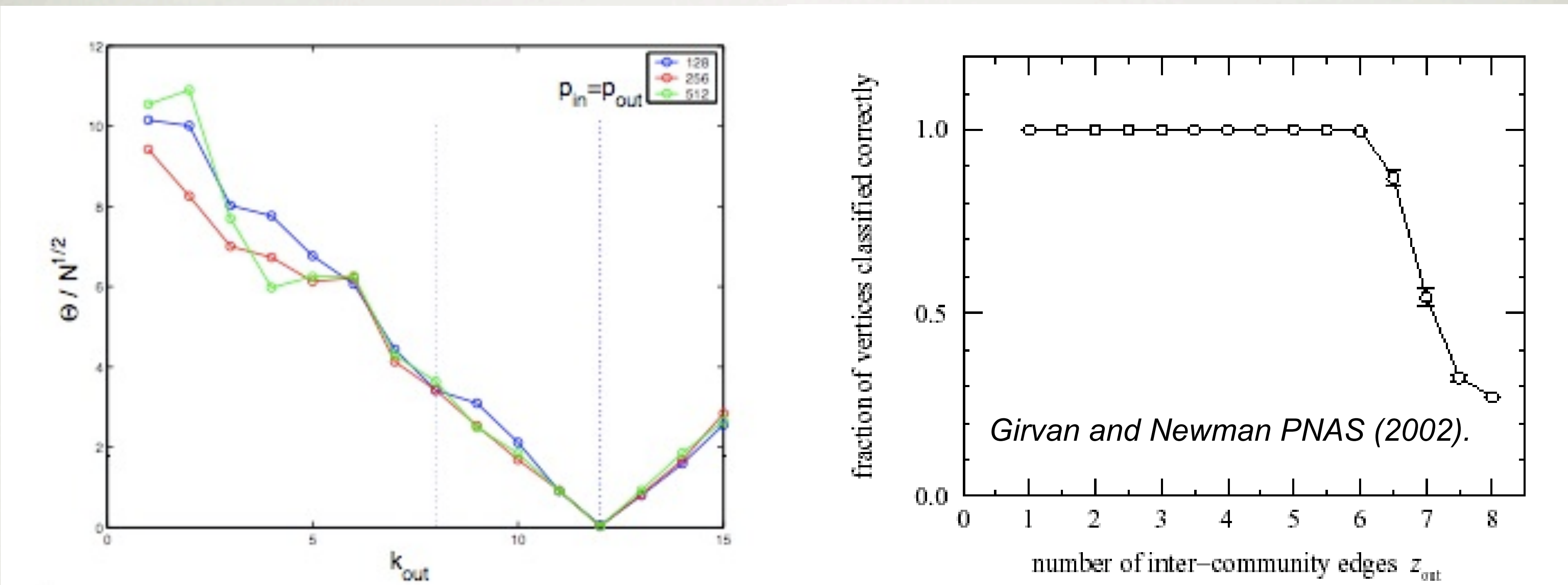
- the same degree sequence of g
- the same number A(q,q′) of links between nodes of type q and q′ (q, q′=1, ..., Q)

$$\Sigma_\kappa = \frac{1}{N}\log(Z_\kappa)\Big|_{h=0}$$

Probability of link i-j $p_{ij} = \frac{\partial \log(Z_\kappa)}{\partial h_{ij}}\Big|_{h=0}$

Benchmarks

- 4 communities with k=16 links/node k_{out} outside community



- features more evident in larger networks ($\Theta \propto \sqrt{n}$)
- even communities not detectable are relevant

Add Health (friendship)

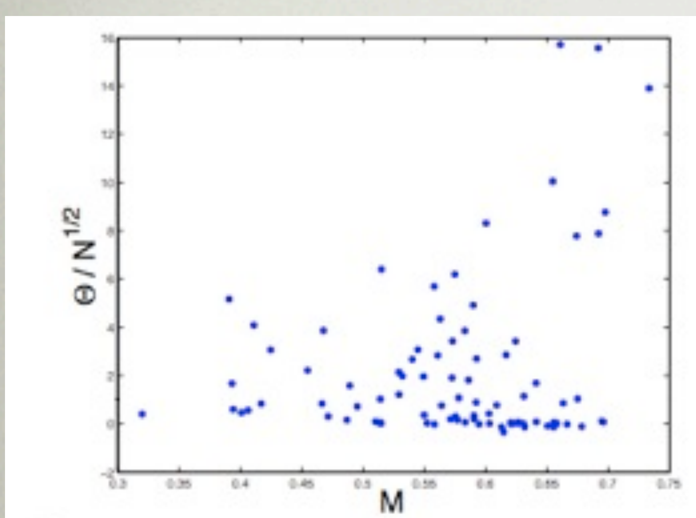
modularity and diversity

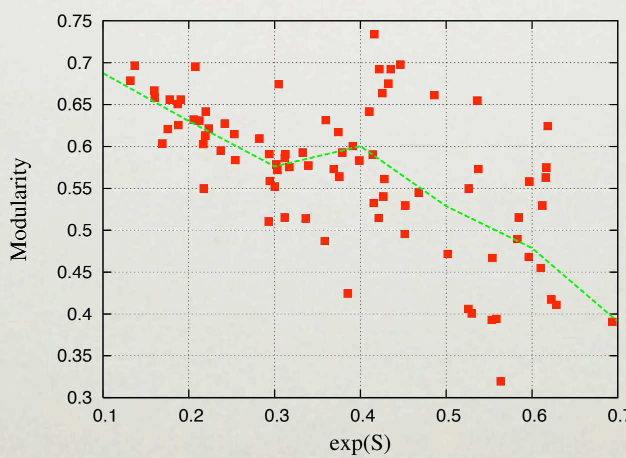
- Modularity

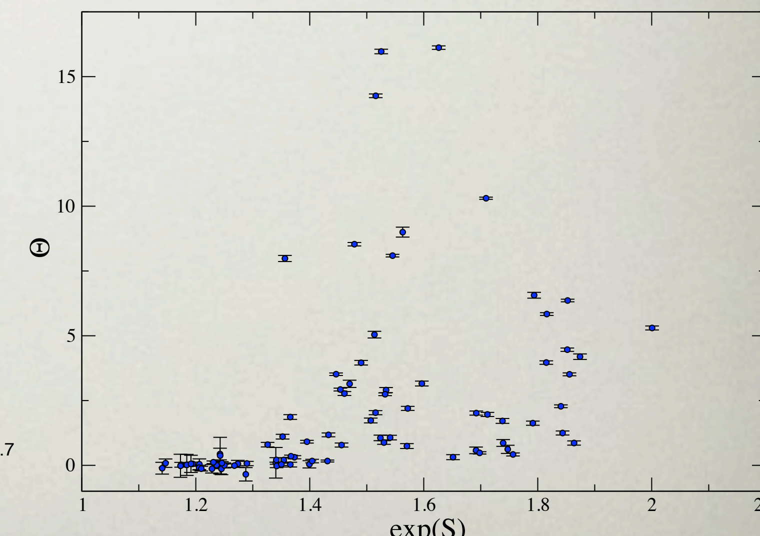
$$M = \sum_{q=1}^{Q} \left[\frac{l_q}{L} - \left(\frac{k_q}{2L} \right)^2 \right]$$

Diversity

$$S = -\sum_{q=1}^{Q} x_q \log x_q$$

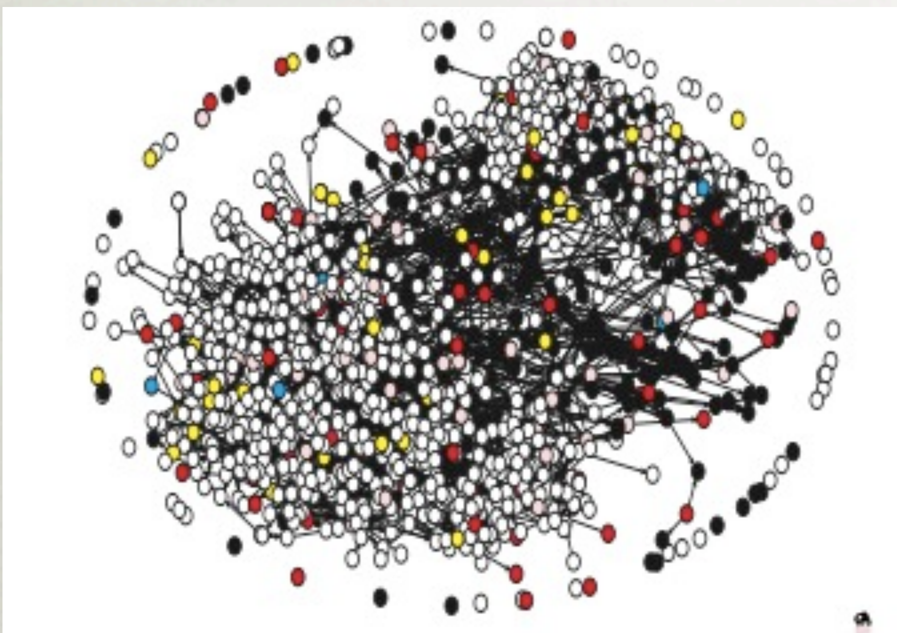


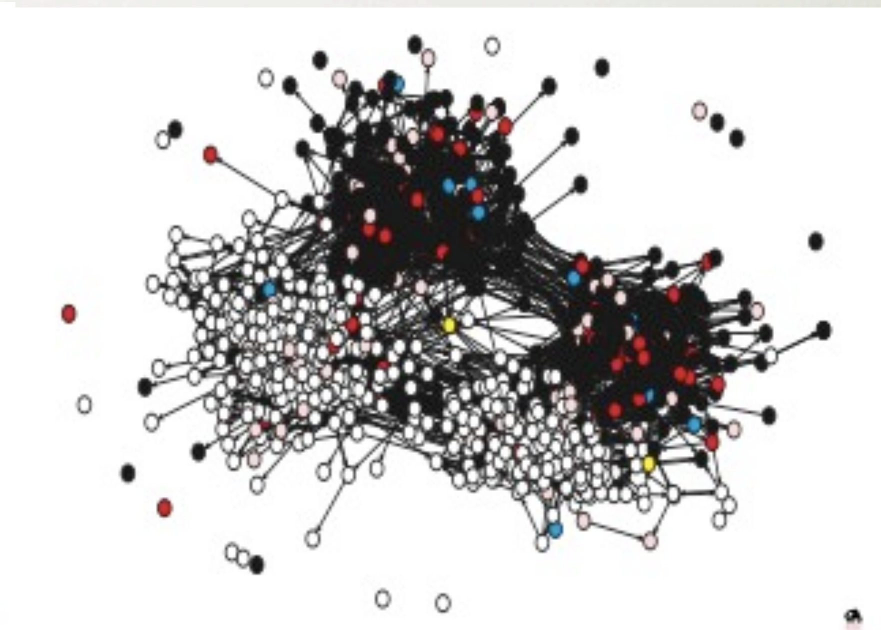




An independent measure

- Two schools with similar n, M, S





$N = 1461,\ M = 0.64,\ S = 0.41,$
$\Omega/\sqrt{N} = 1.69$

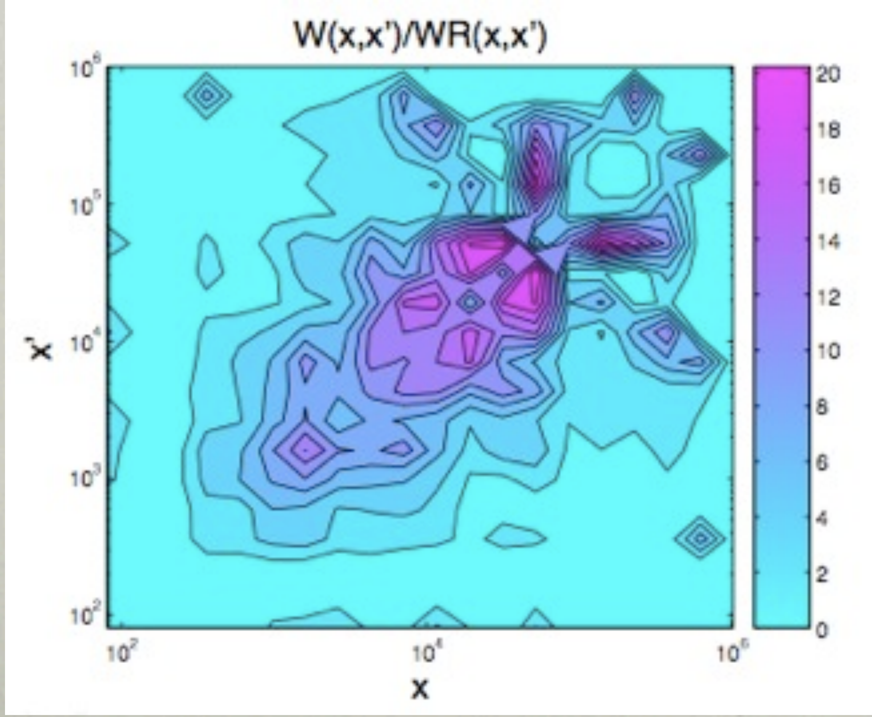
$N = 1147,\ M = 0.66,\ S = 0.48,$
$\Omega/\sqrt{N} = 15.71$

Is abundance relevant in p-p interaction net?

Protein-protein interaction network of Saccharomyces Cerevisiae. N=1740, L=4185 (Maslov, Ispolatov 2007)

50-1000000 molecules/cell: x_i=log(abundance protein i)

x_i not correlated with degree (R=0.13) or clustering (R=0.005)



$$\Theta \simeq 22, \qquad \Theta/\sqrt{N} \simeq 0.52 \quad .$$

$$P\{\Theta > 2.7\} \leq 0.01$$

$$p_{i,j} = \frac{\theta_i \theta_j W(x_i, x_j)}{1 + \theta_i \theta_j W(x_i, x_j)}$$

Feature = position

$\Sigma(g,B)$ = log Number of networks with

- the same degree sequence of g
- the same number B(d) of links between nodes at distance d

$$\Sigma_\kappa = \frac{1}{N}\log(Z_\kappa)\Big|_{h=0}$$

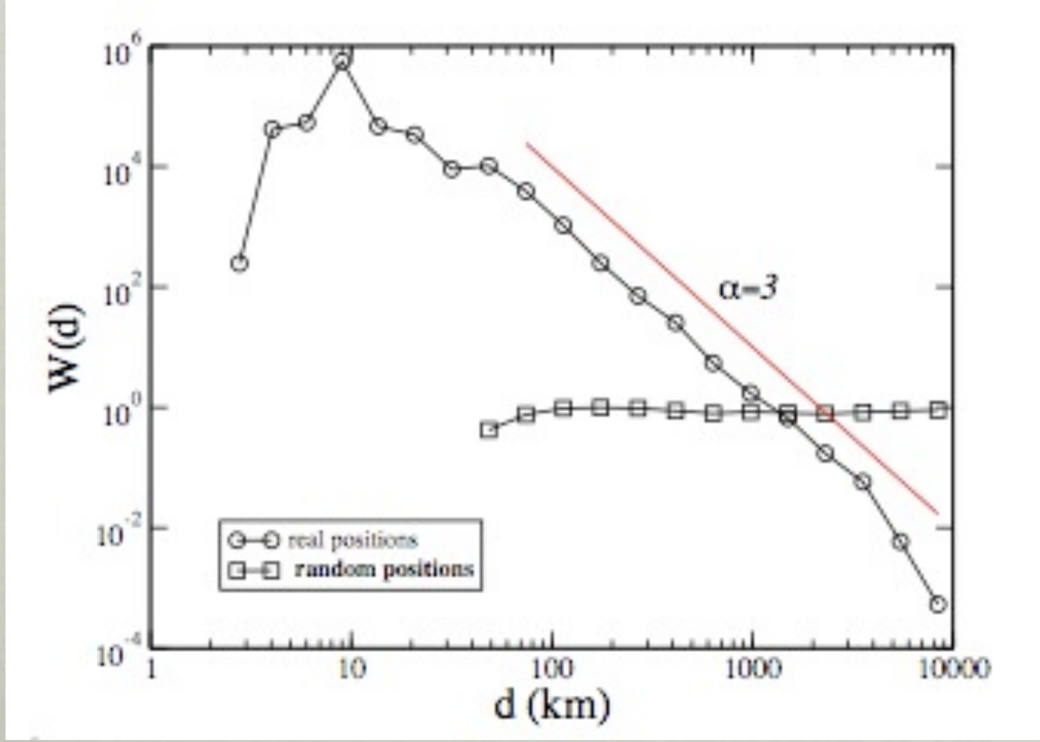
Probability of link i-j $\quad p_{ij} = \dfrac{\theta_i\theta_j W(d_{ij})}{1+\theta_i\theta_j W(d_{ij})}$

Is geographic location relevant for airport net?

IATA data N=675 airports, L=3253 flight connections (Colizza et al, Nature Phys. 2007)

$\Theta \simeq 1.1 \cdot 10^3, \qquad \Theta/\sqrt{N} \simeq 42$

$$p_{ij} = \frac{\theta_i \theta_j W(d_{ij})}{1 + \theta_i \theta_j W(d_{ij})}$$



Costs of flights longer than R

$$C(R) \propto \int_R^\infty r^2 W(r) dr \sim R^{3-\alpha}$$

Optimal $\alpha = 2$ (Kleinberg, 2000)

Competitive market $\alpha \geq 3$

(but see P. De Los Rios arxiv 2009)

Conclusion

- Inferring properties of underlying network
  - distinguish causes of homophily (choice and opportunity)
  - measuring the relevance of features
    - universal indicator, non-reducible to known measures
    - extensions: other features/directed networks
    - uncovering hidden statistical regularities relevant for network stability or formation