

Social Network Analysis

A crash course @ UPF

Dino Pedreschi



ISTI-CNR & Università di Pisa

<http://kdd.isti.cnr.it>



Complex (Social) Networks

- Big graph data and social, information, biological and technological networks
- The architecture of complexity and how real networks differ from random networks:
 - node degree and long tails,
 - social distance and small worlds,
 - clustering and triadic closure.
- Comparing real networks and random graphs.
- The main models of network science: small world and preferential attachment.



Complex (Social) Networks

- Strong and weak ties, community structure and long-range bridges.
- Robustness of networks to failures and attacks.
- Cascades and spreading. Network models for diffusion and epidemics. The strength of weak ties for the diffusion of information. The strength of strong ties for the diffusion of innovation.
- Practical network analytics with Cytoscape and Gephi.
- Simulation of network processes with NetLogo.



Complex (Social) Networks

- Textbooks
 - Albert-Laszlo Barabasi. *Network Science* (2016)
 - <http://barabasi.com/book/network-science>
 - David Easley, Jon Kleinberg: *Networks, Crowds, and Markets* (2010)
 - <http://www.cs.cornell.edu/home/kleinber/networks-book/>
- Network Analytics Software (open):
 - Cytoscape: <http://www.cytoscape.org/>
 - Gephi: <http://gephi.github.io/>
- Network Data Repository
 - <http://networkrepository.com/>
- Simulation of network models: NetLogo

Part 3

- Robustness
- Cascades
- Models of diffusion and spreading
- Empirical studies
- Research highlights at KDD LAB Pisa

Network robustness

A SIMPLE STORY (3):



ROBUSTNESS IN COMPLEX SYSTEMS

Complex systems maintain their basic functions even under errors and failures

cell → mutations

There are uncountable number of mutations and other errors in our cells, yet, we do not notice their consequences.

Internet → router breakdowns

At any moment hundreds of routers on the internet are broken, yet, the internet as a whole does not lose its functionality.

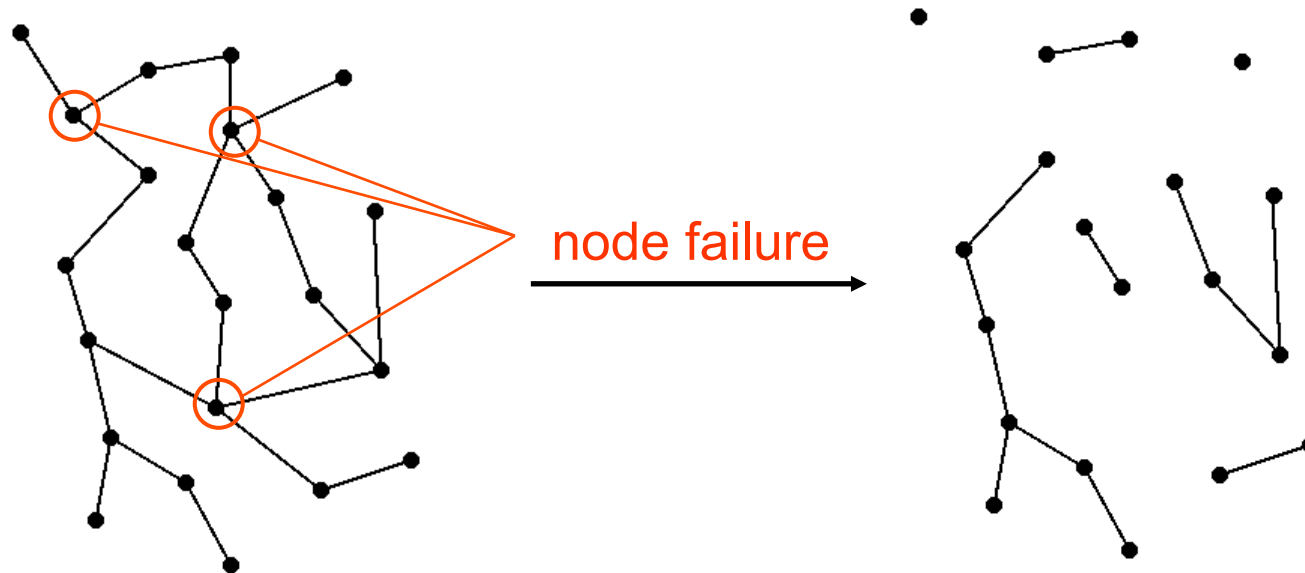
Where does robustness come from?

There are feedback loops in most complex systems that keep tabs on the components and the system's 'health'.

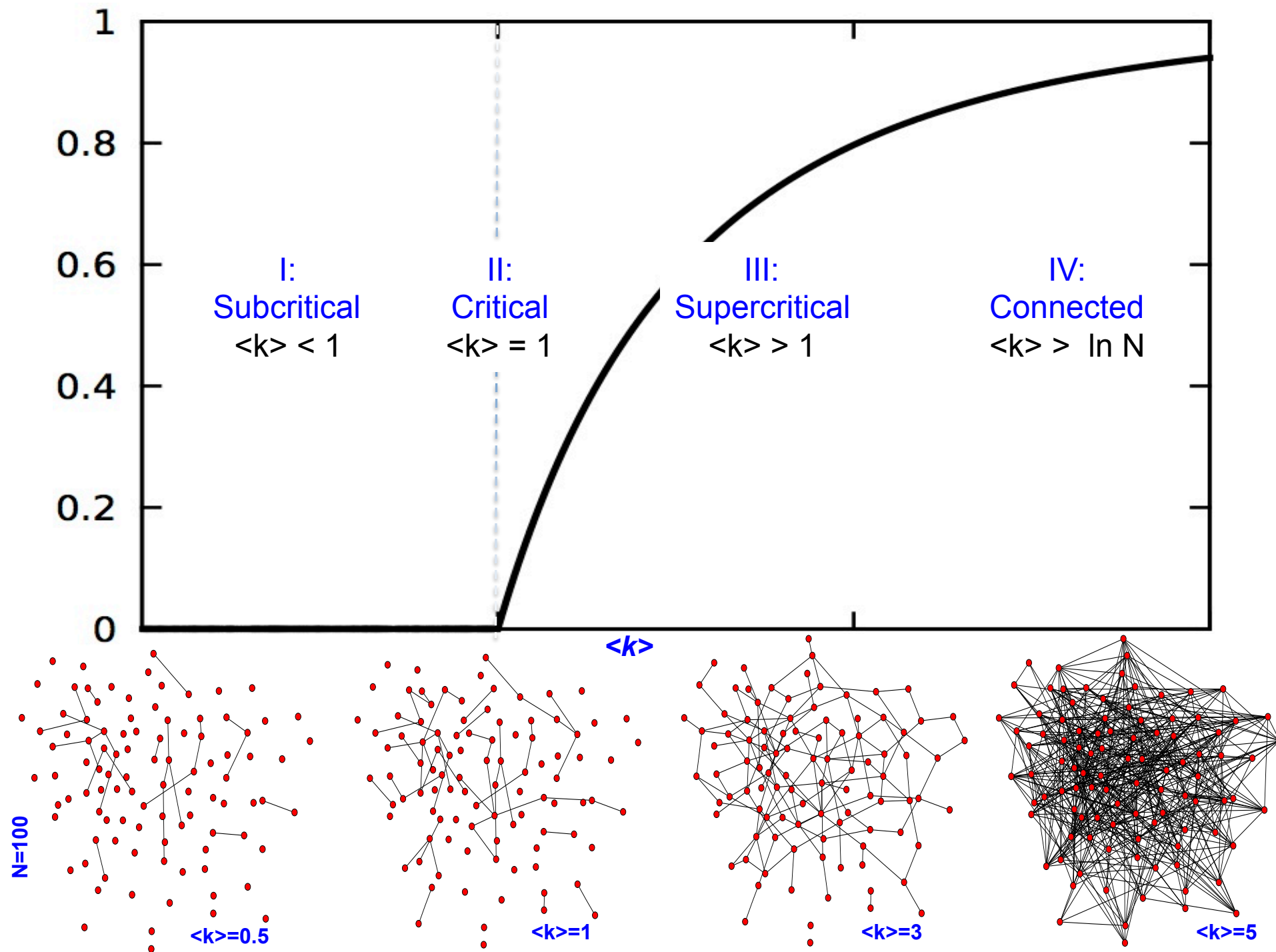
Could the network structure affect a system's robustness?

ROBUSTNESS

Could the network structure affect a system's robustness?

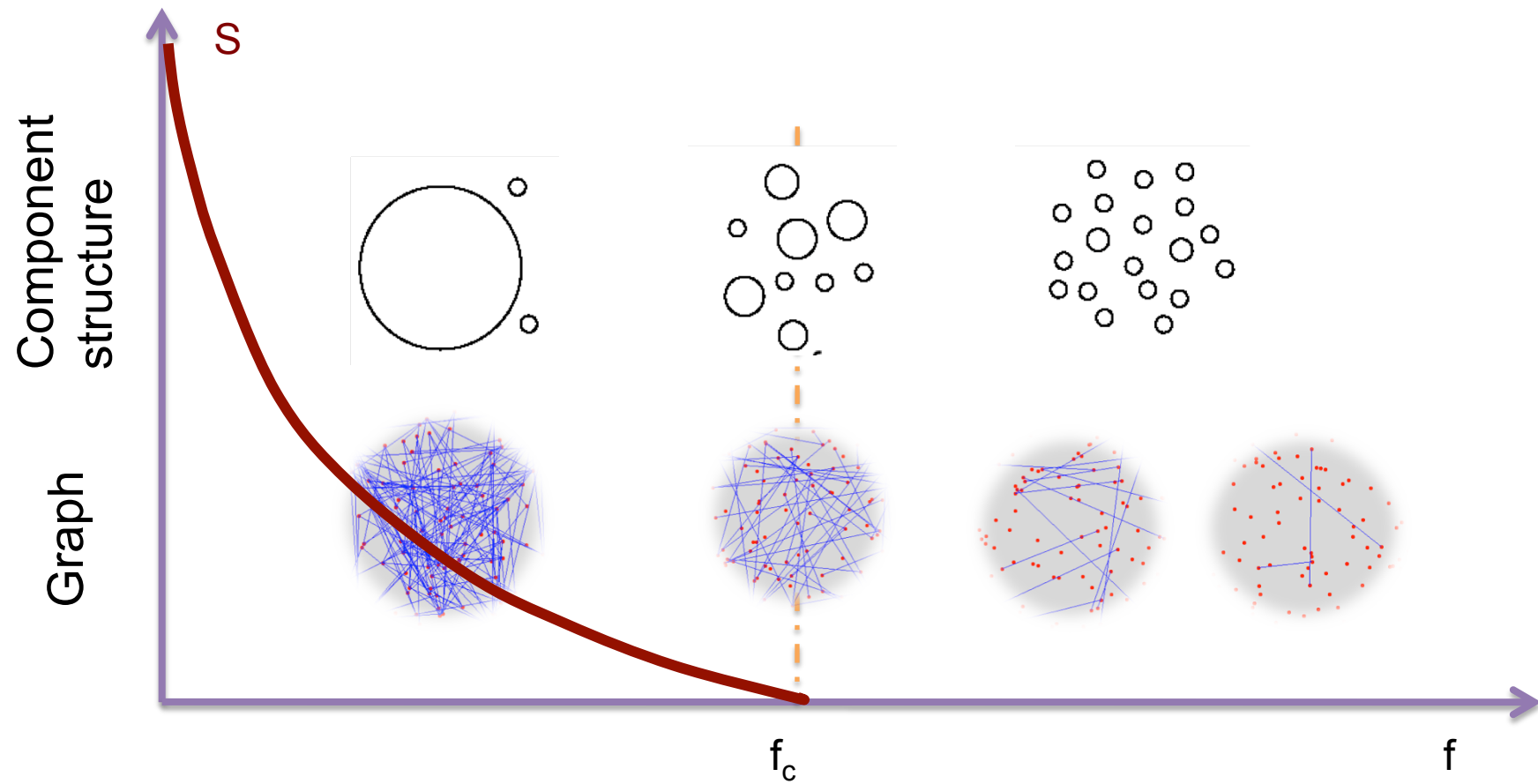


How do we describe in quantitative terms the breakdown of a network under node or link removal?
~percolation theory~



Damage is modeled as an inverse percolation process

f = fraction of removed nodes



(Inverse Percolation phase transition)

ROBUSTNESS: OF SCALE-FREE NETWORKS

The interest in the robustness problem has three origins:

- Robustness of complex systems is an important problem in many areas
- Many real networks are not regular, but have a scale-free topology
- *In scale-free networks the scenario described above is not valid*

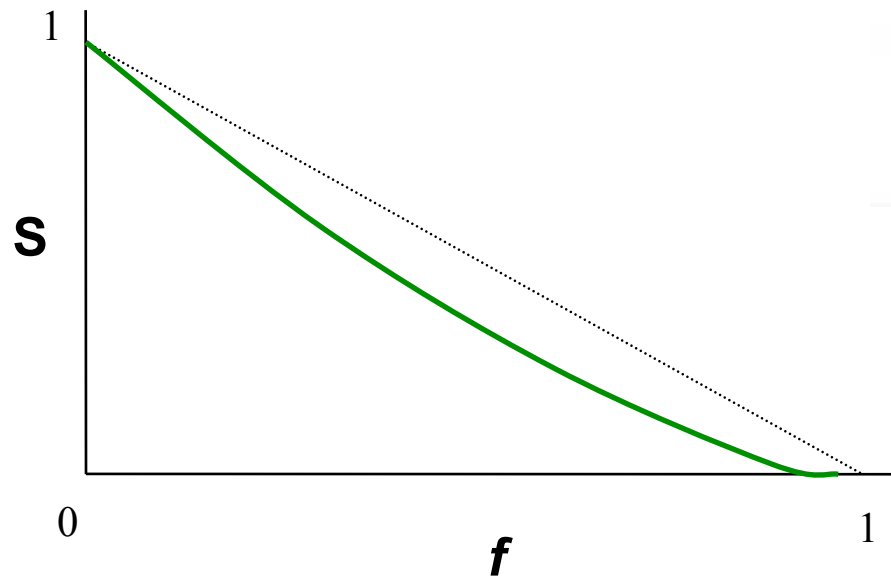
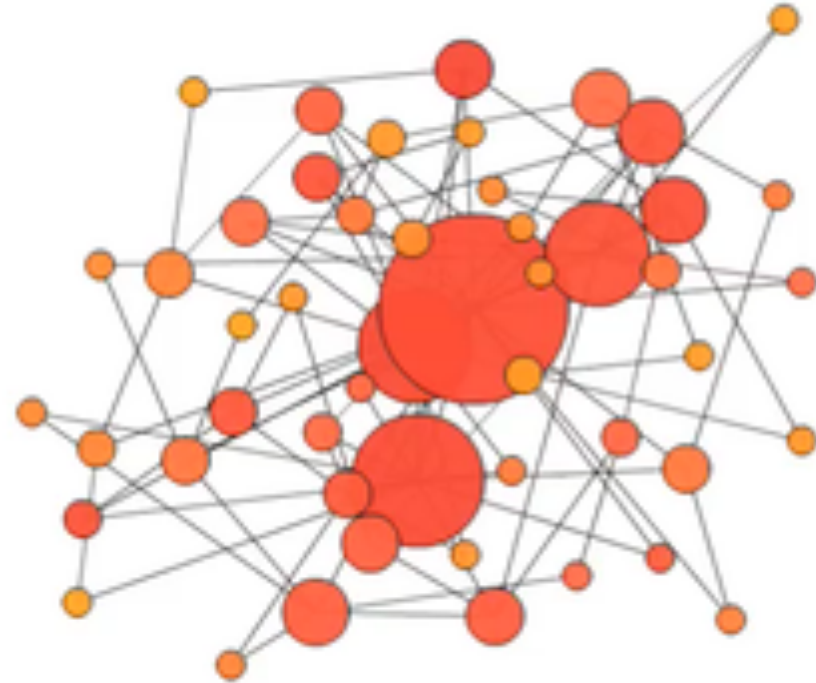
Albert, Jeong, Barabási, *Nature* **406** 378 (2000)

ROBUSTNESS OF SCALE-FREE NETWORKS

Scale-free networks do not appear to break apart under random failures.

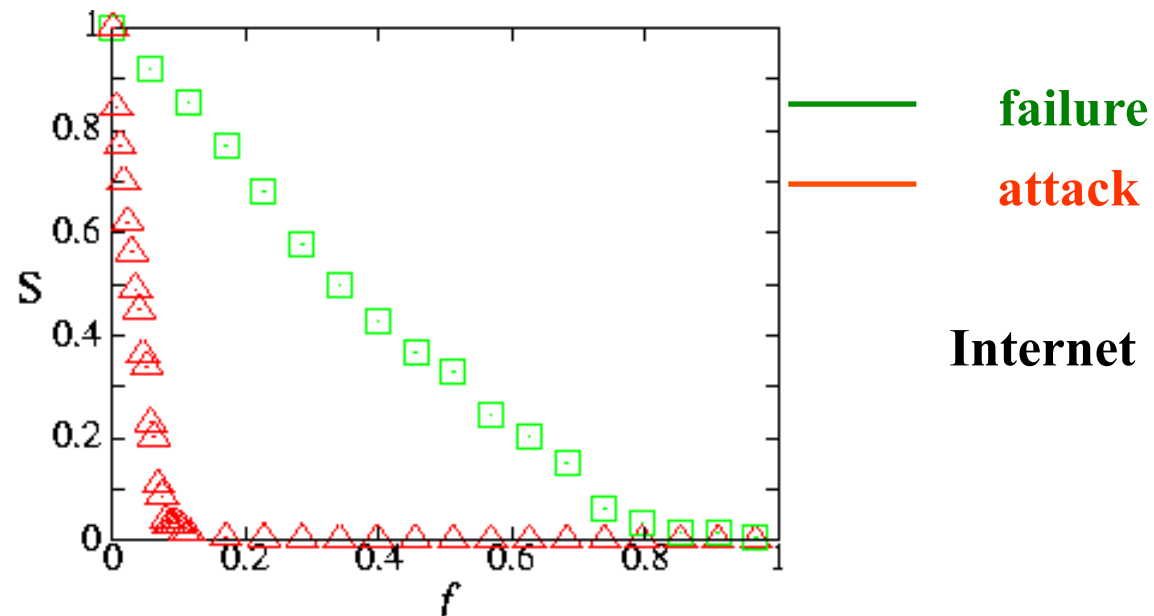
Reason: the hubs.

The likelihood of removing a hub is small.



Albert, Jeong, Barabási, *Nature* **406** 378 (2000)

INTERNET'S ROBUSTNESS TO RANDOM FAILURES



R. Albert, H. Jeong, A.L. Barabasi, *Nature* **406** 378 (2000)

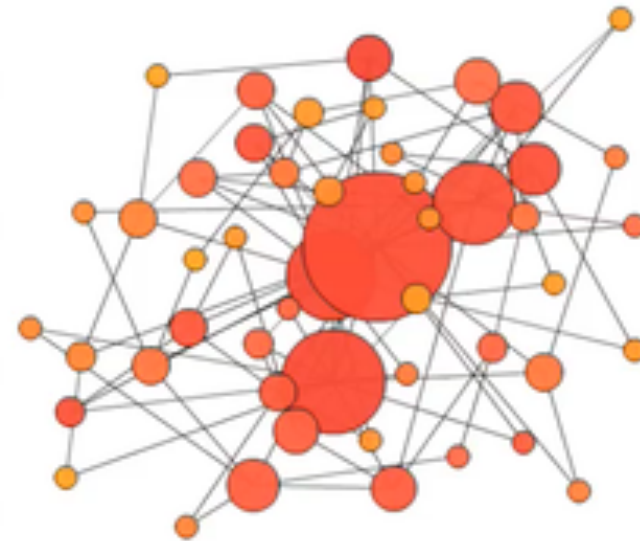
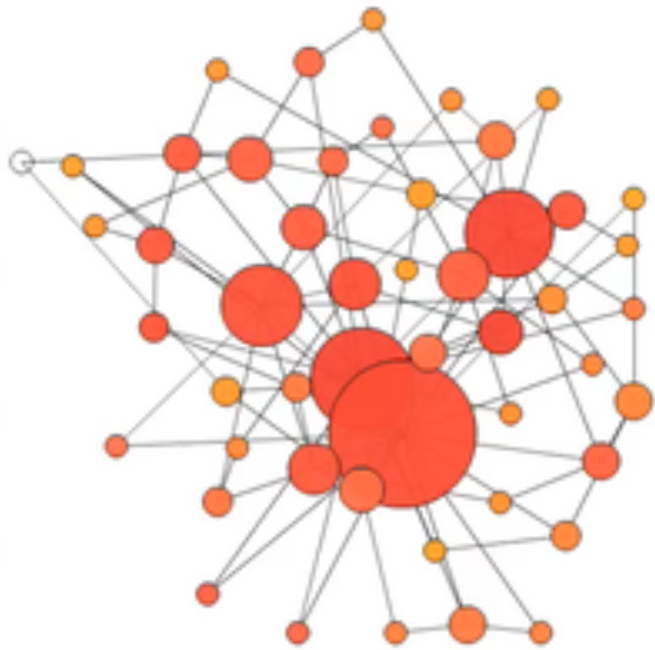
$$f_c = 1 - \frac{1}{\kappa - 1}$$

Internet: Router level map, $N=228,263$; $\gamma=2.1\pm0.1$; $\kappa=28 \rightarrow f_c=0.962$

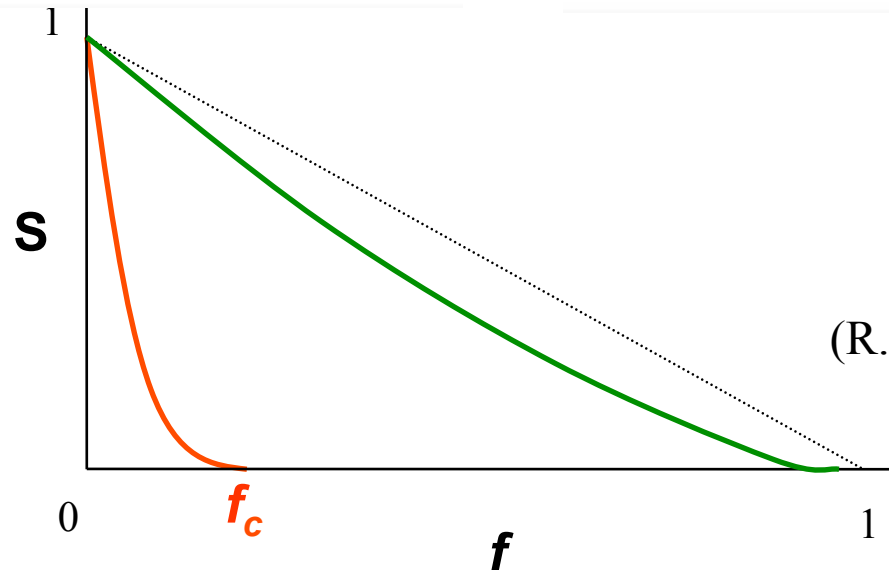
AS level map, $N=11,164$; $\gamma=2.1\pm0.1$; $\kappa=264 \rightarrow f_c=0.996$

Internet parameters: Pastor-Satorras & Vespignani, *Evolution and Structure of the Internet*: Table 4.1 & 4.4

Achilles' Heel of scale-free networks



Attacks

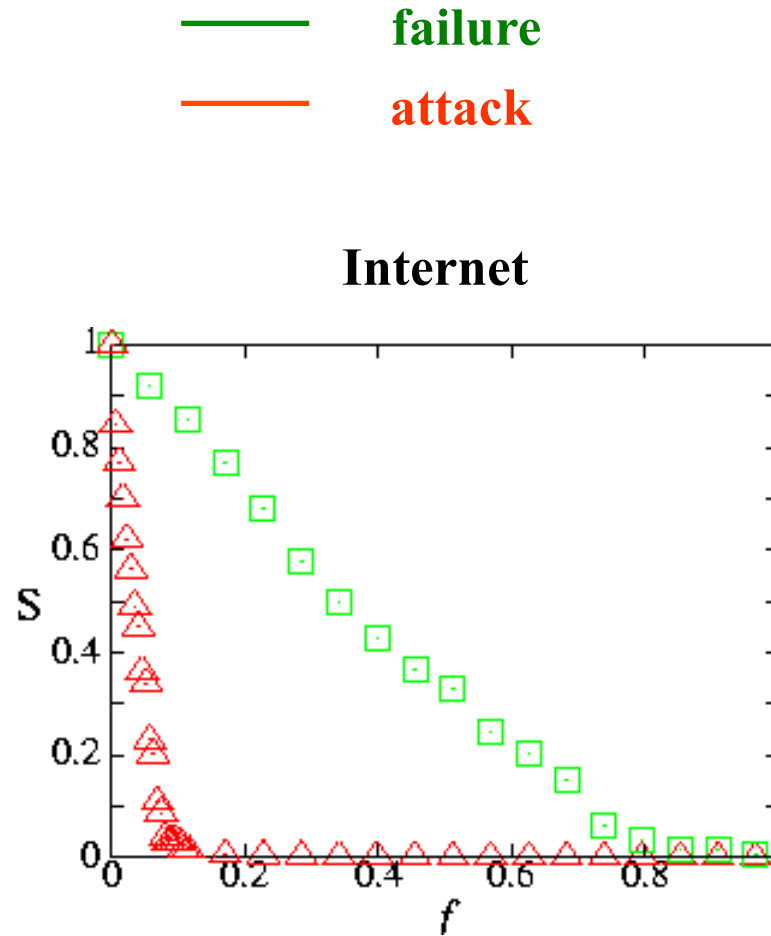


Failures

$$\gamma \leq 3 : f_c = 1$$

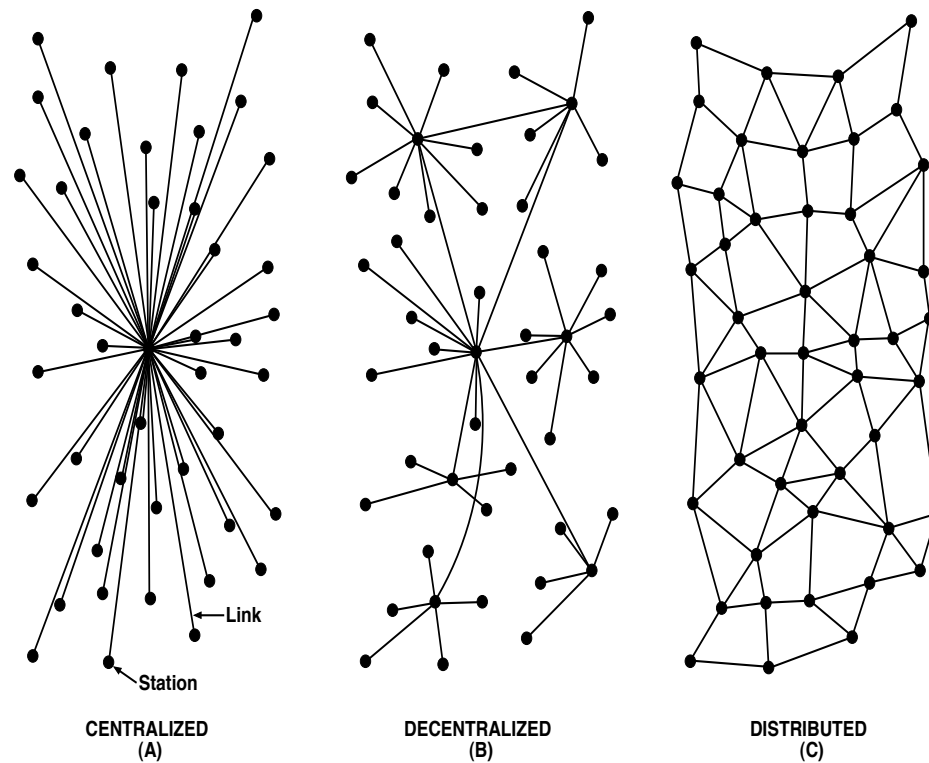
(R. Cohen et al PRL, 2000)

Achilles' Heel of complex networks



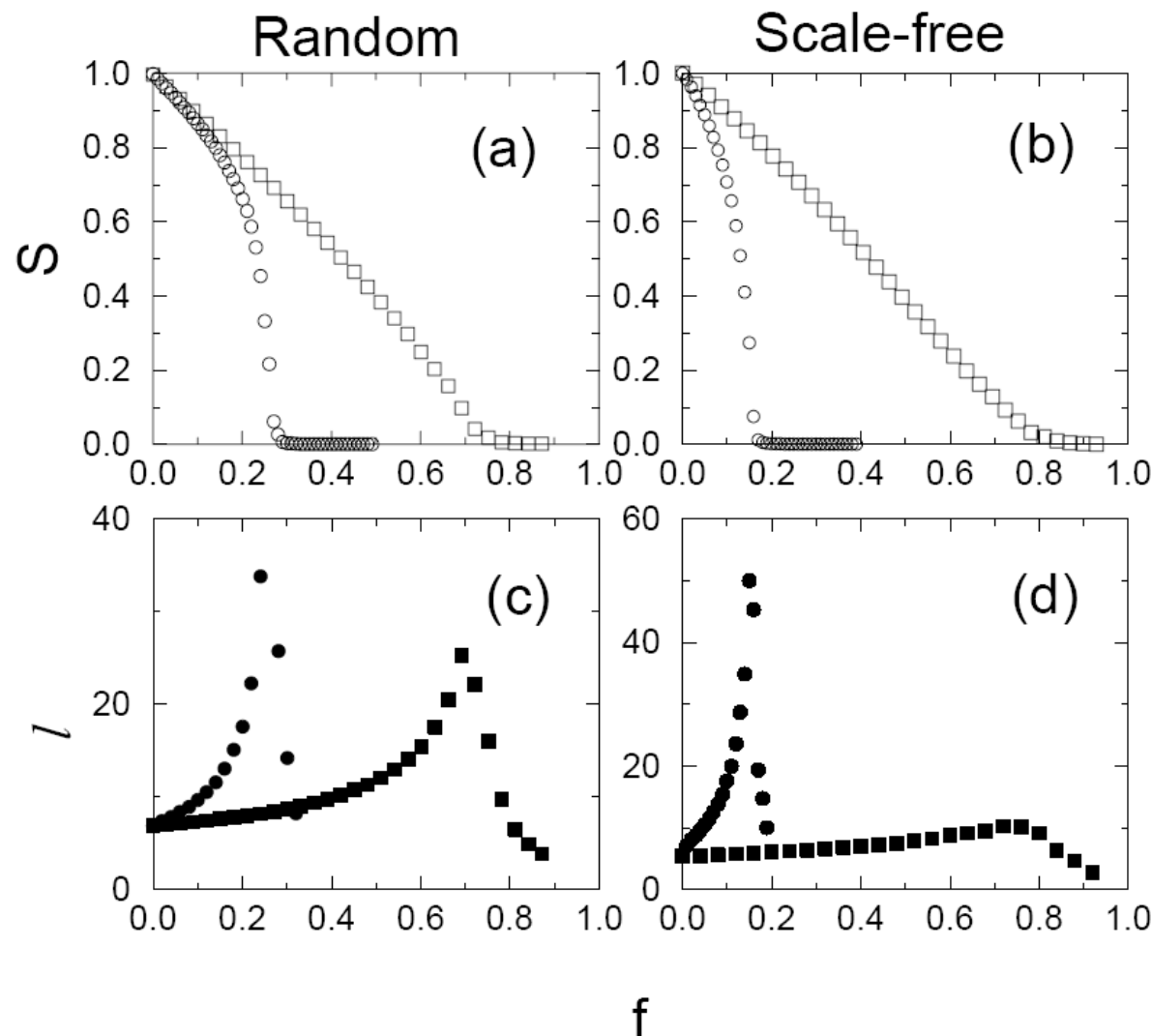
R. Albert, H. Jeong, A.L. Barabasi, *Nature* **406** 378 (2000)

Historical Detour: Paul Baran and Internet



1958

Scale-free networks are more error tolerant, but also more vulnerable to attacks



- squares: random failure
- circles: targeted attack

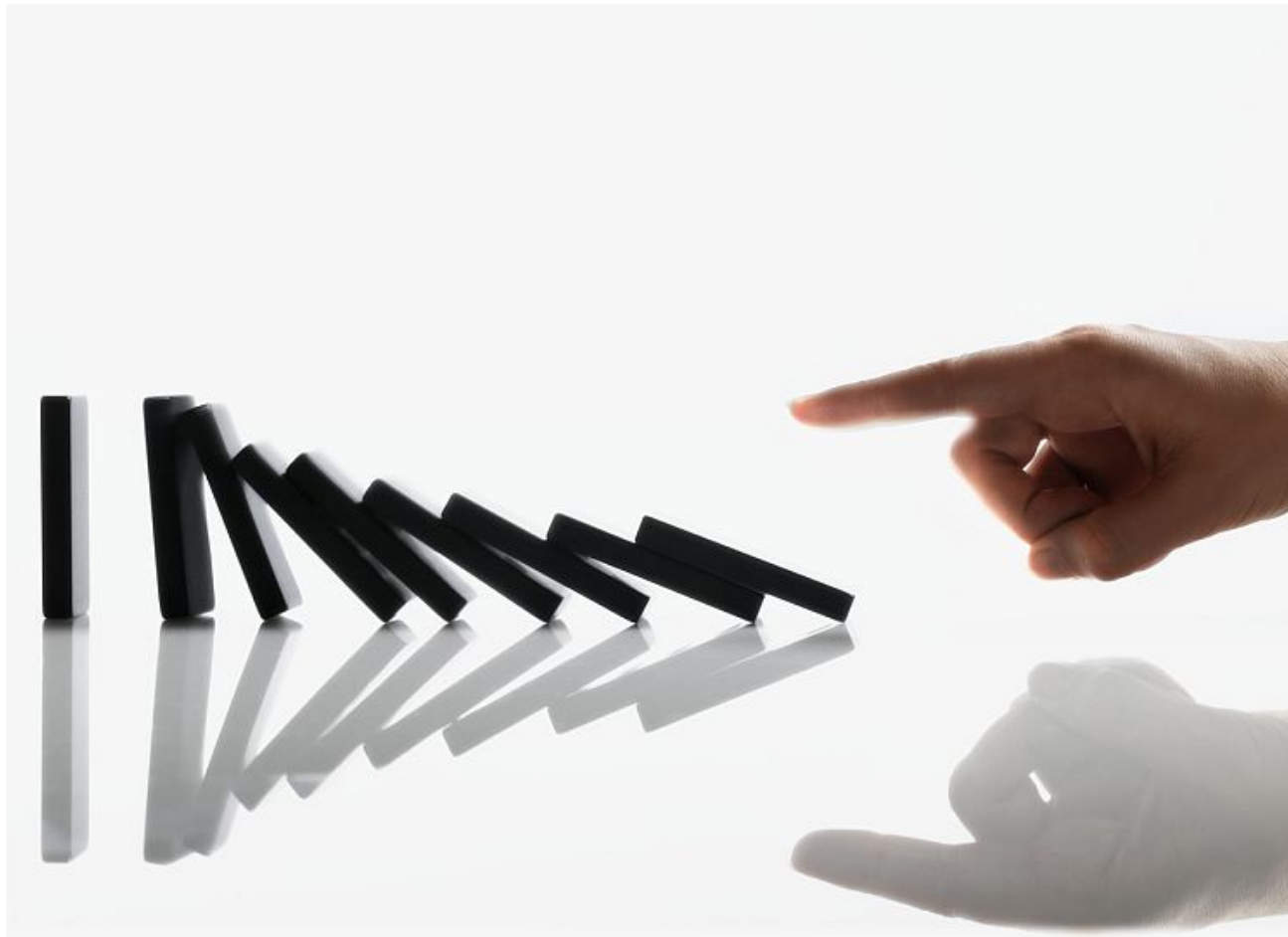
Failures: little effect on the integrity of the network.

Attacks: fast breakdown

Cascades

Cascades

Potentially large events triggered by small initial shocks



- **Information cascades**
social and economic systems
diffusion of innovations
- **Cascading failures**
infrastructural networks
complex organizations

Cascading Failures in Nature and Technology

Blackout



Earthquake



Avalanche



Flows of physical quantities

- congestions
- instabilities
- Overloads

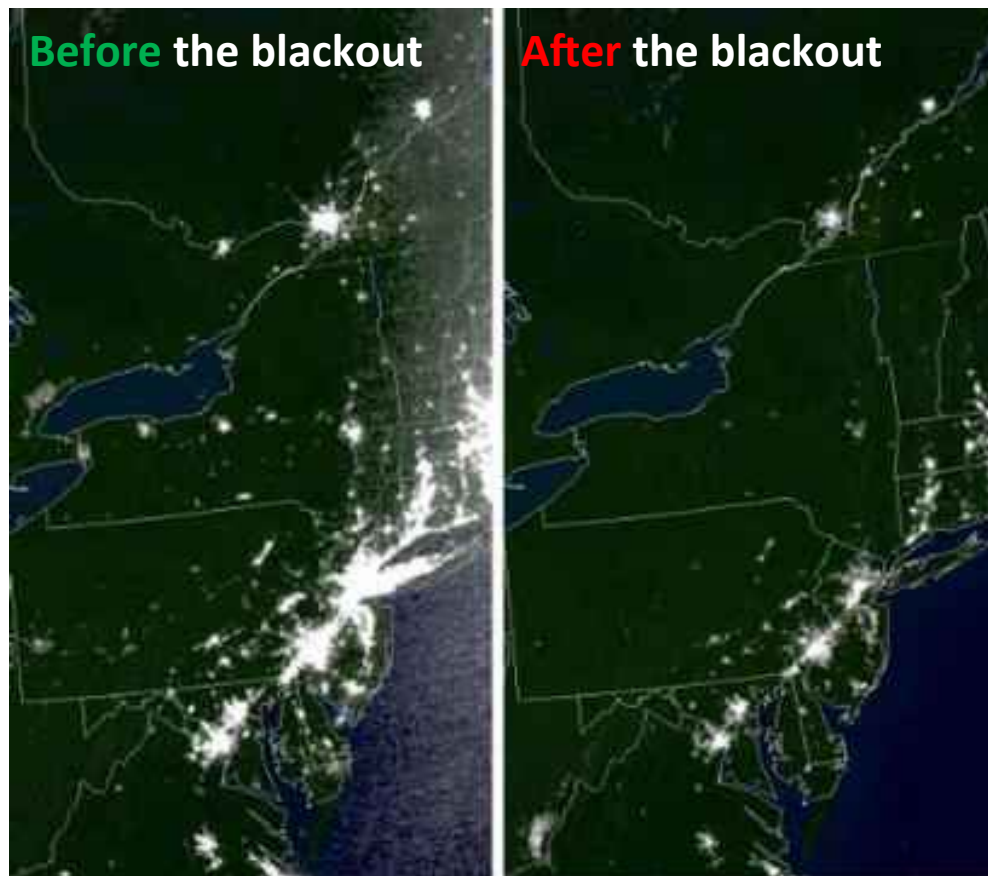
Cascades depend on

- Structure of the network
- Properties of the flow
- Properties of the net elements
- Breakdown mechanism

Northeast Blackout of 2003

Origin

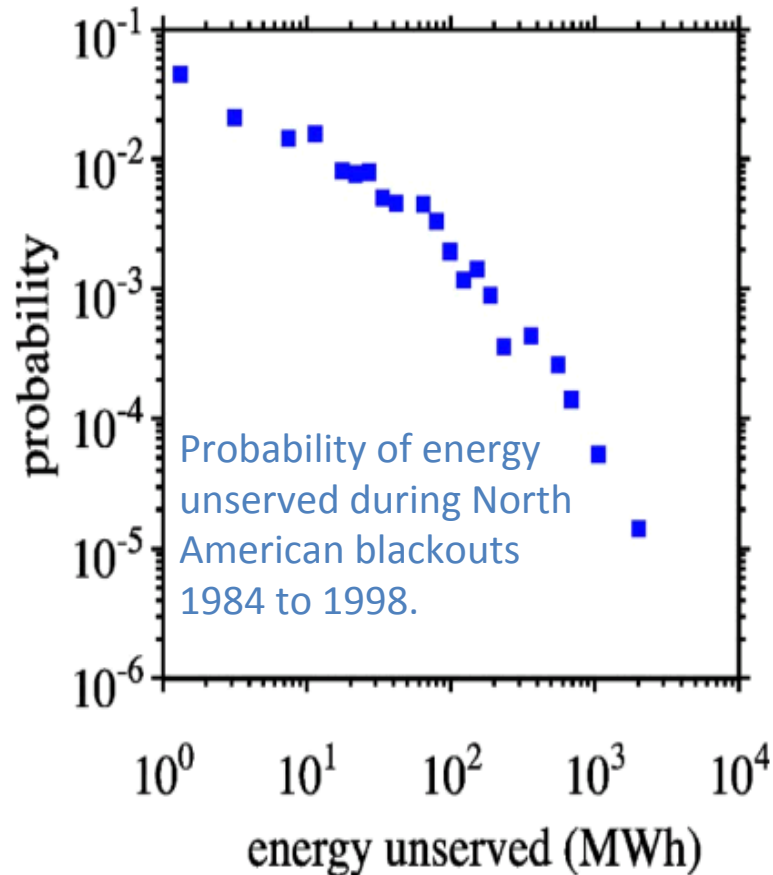
A 3,500 MW power surge (towards Ontario) affected the transmission grid at 4:10:39 p.m. EDT. (Aug-14-2003)



Consequences

More than 508 generating units at 265 power plants shut down during the outage. In the minutes before the event, the NYISO-managed power system was carrying 28,700 MW of load. At the height of the outage, the load had dropped to 5,716 MW, a loss of 80%.

Cascades Size Distribution of Blackouts



Unserved energy/power magnitude (S) distribution

$$P(S) \sim S^{-\alpha}, 1 < \alpha < 2$$

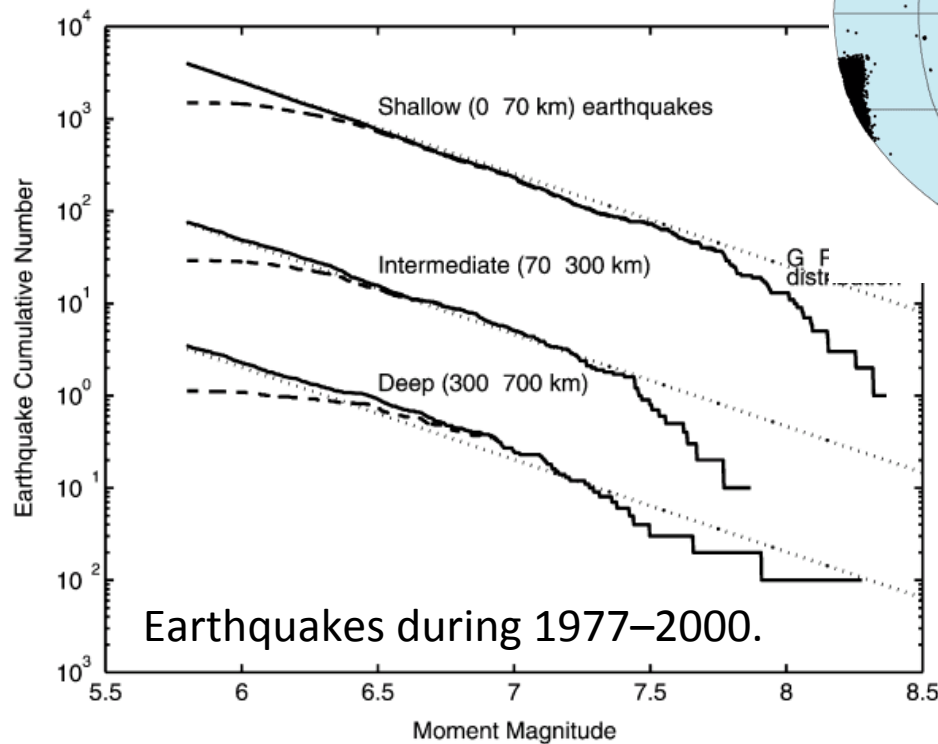
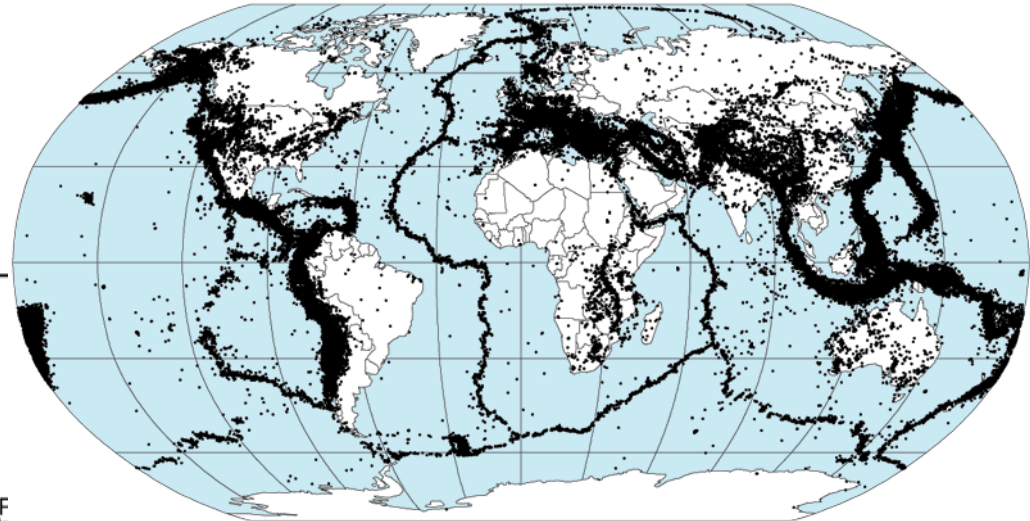
Source	Exponent	Quantity
North America	2.0	Power
Sweden	1.6	Energy
Norway	1.7	Power
New Zealand	1.6	Energy
China	1.8	Energy

I. Dobson, B. A. Carreras, V. E. Lynch, D. E. Newman, *CHAOS* 17, 026103 (2007)

Cascades Size Distribution of Earthquakes

Preliminary Determination of Epicenters

358,214 Events, 1963 - 1998



Earthquake size S distribution

$$P(S) \sim S^{-\alpha}, \alpha \approx 1.67$$

Short Summary of Models: Universality

Models	Networks	Exponents
Failure Prorogation Model	ER	1.5
Overload Model	Complete Graph	1.5
BTW Sandpile Model	ER/SF	1.5 (ER) $\gamma/(\gamma - 1)$ (SF)
Branching Process Model	ER/SF	1.5 (ER) $\gamma/(\gamma - 1)$ (SF)

Universal for homogenous networks

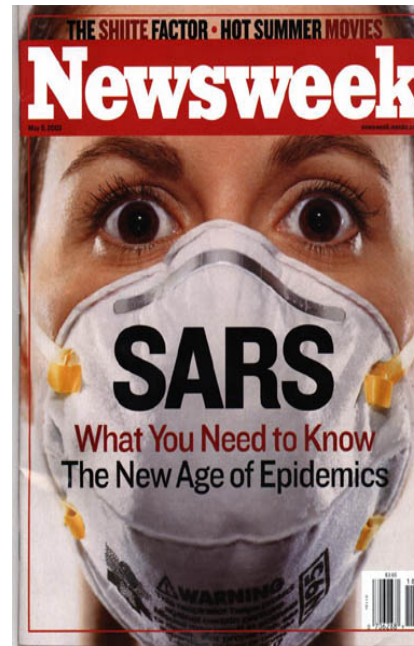
$$P(S) \sim S^{-3/2}$$

Same exponent for percolation too
(random failure, attacking, etc.)

Epidemics and spreading

Epidemic spreading – Why?

Why is the spreading process important?



“Epidemic”

Epi + demos

upon

people



Biological:

Airborne diseases (flu, SARS, ...)

- Venereal diseases (HIV, ...)
- Other infectious diseases including some cancers (HPV, ...)
- Parasites (bedbugs, malaria, ...)

Digital:

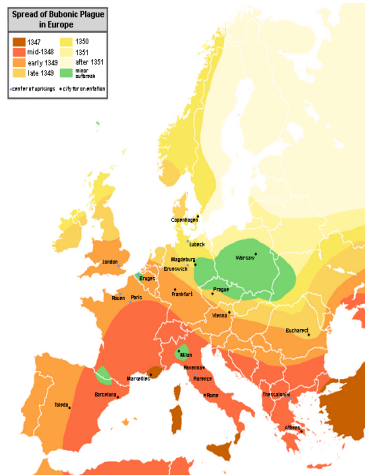
- Computer viruses, worms
- Mobile phone viruses

Conceptual/Intellectual:

- Diffusion of innovations
- Rumors
- Memes
- Business practices

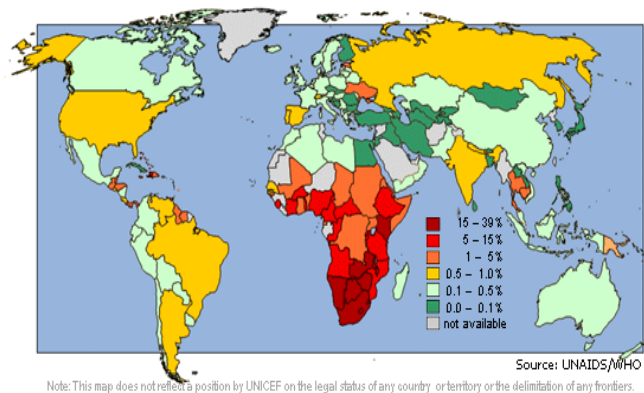
Biological: Notable Epidemic Outbreaks

The Great Plague

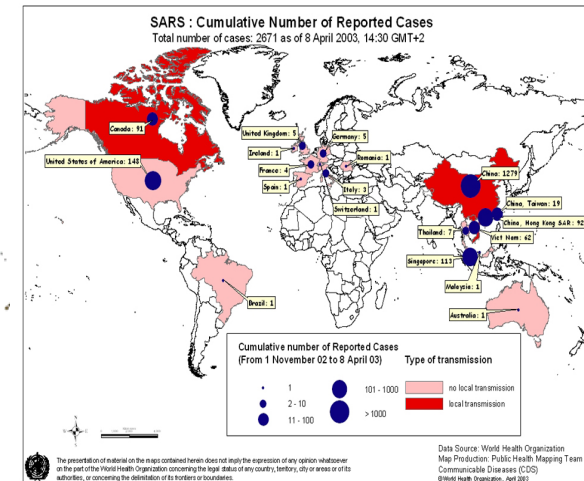


HIV

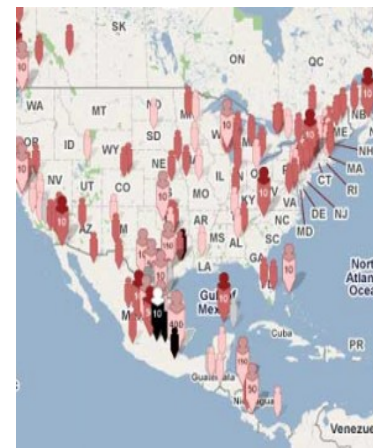
HIV prevalence in adults, end 2001



SARS



1918 Spanish flu



H1N1 flu

Epidemic spreading – Why does it matter now?

High population density



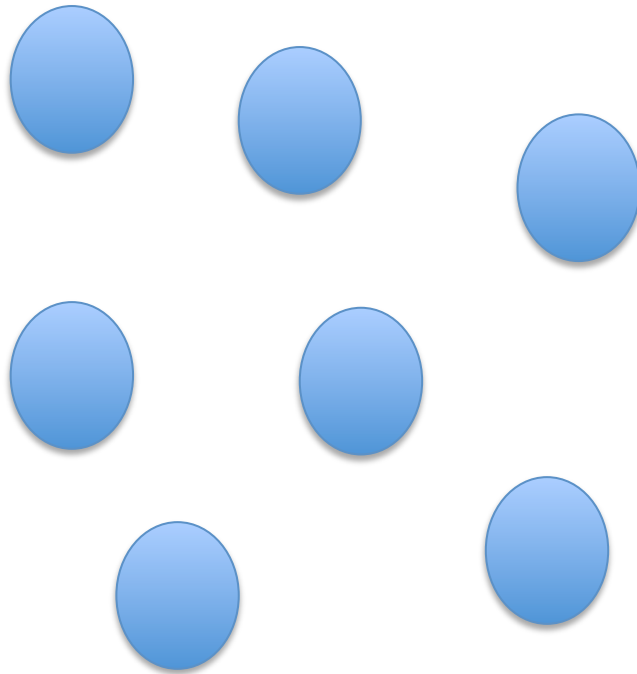
High mobility



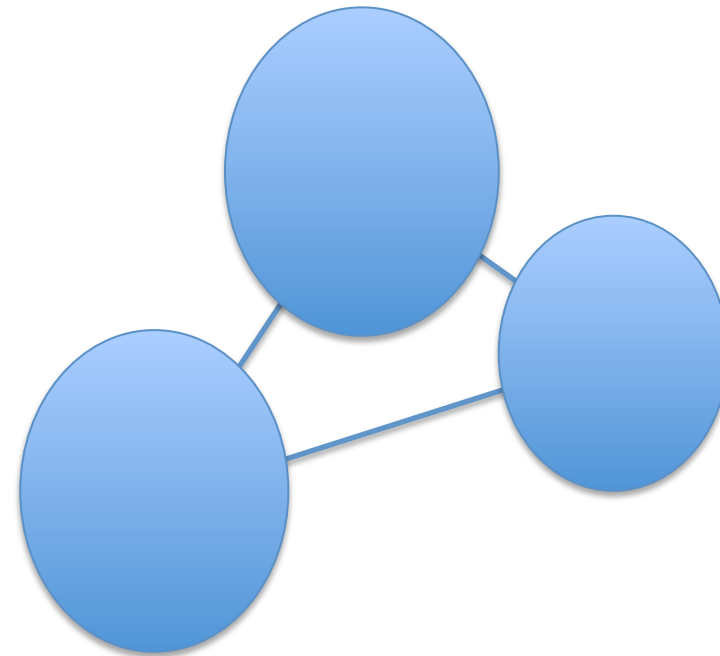
→ perfect conditions for epidemic spreading.

Airline figure: L. Hufnagel et al. *PNAS* **101**, 15124 (2004)

Large population can provide the “fuel”



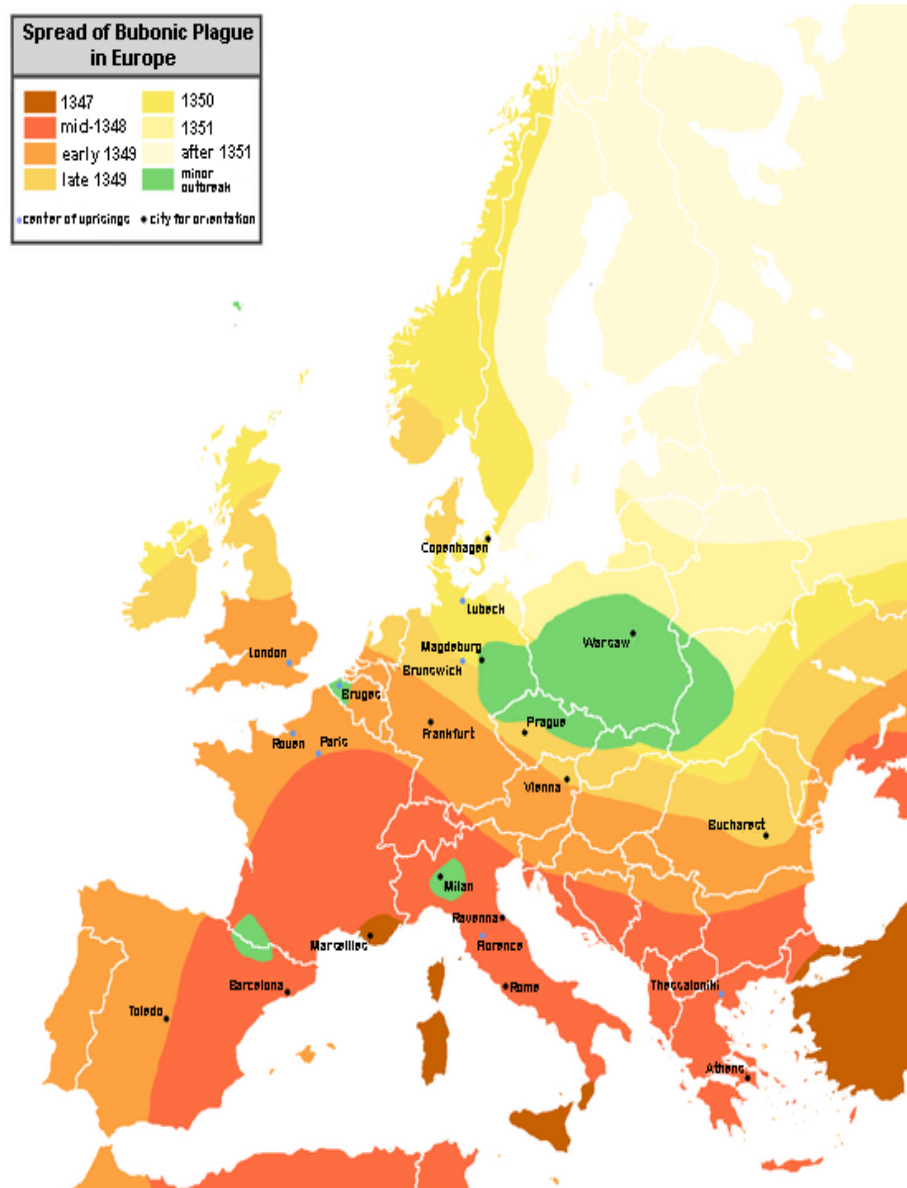
Separate, small population
(hunter-gatherer society, wild animals)



Connected, highly populated areas
(cities)

Human societies have “**crowd diseases**”, which are the consequences of large, interconnected populations (Measles, tuberculosis, smallpox, influenza, common cold, ...)

14th Century – The Great Plague

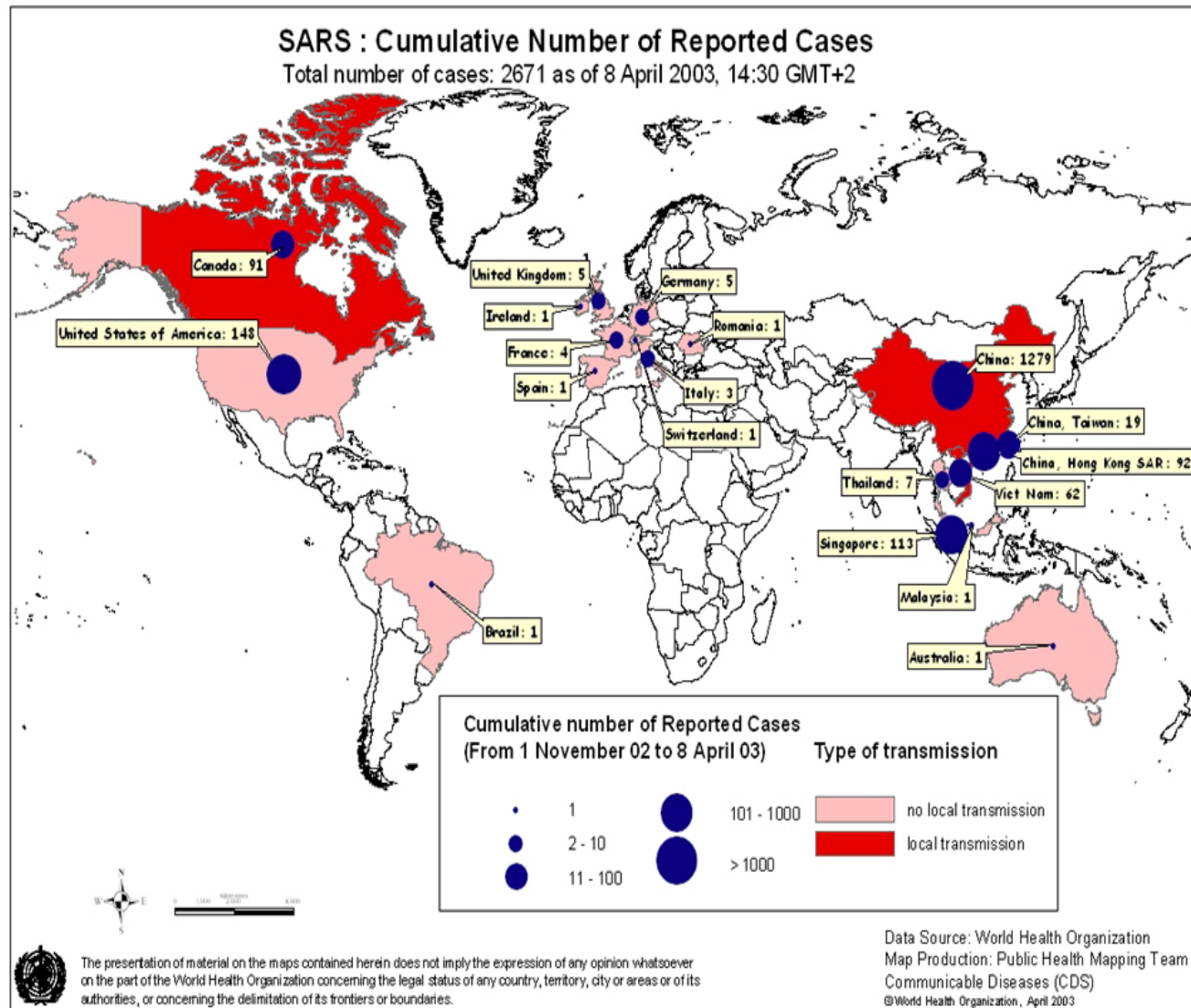


4 years from France to Sweden

Limited by the speed of human travel

http://en.wikipedia.org/wiki/Black_Death
http://de.wikipedia.org/wiki/Schwarzer_Tod

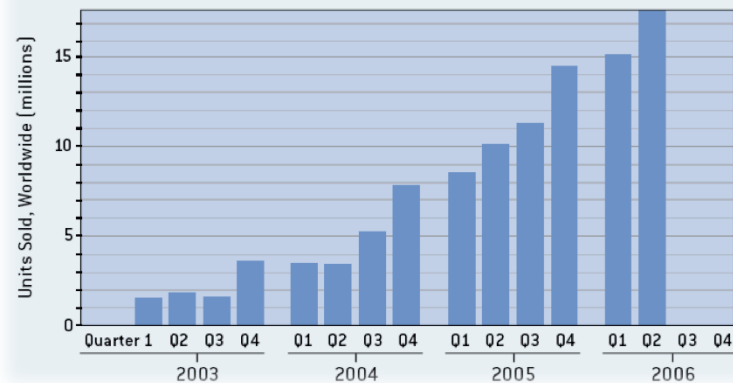
21st Century – SARS



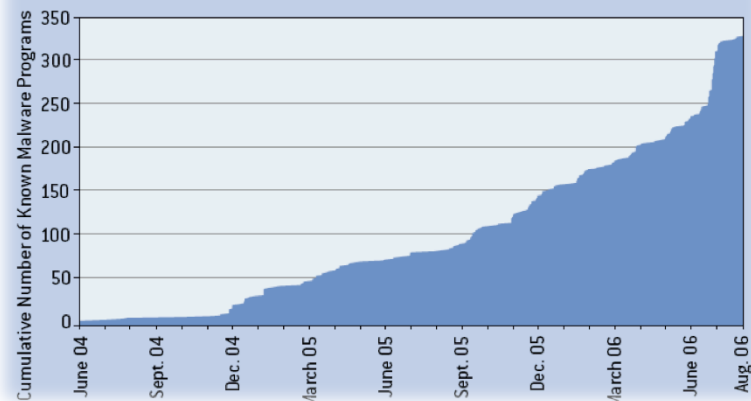
Source: World Health Organization

Computer Viruses, Worms, Mobile Phone Viruses

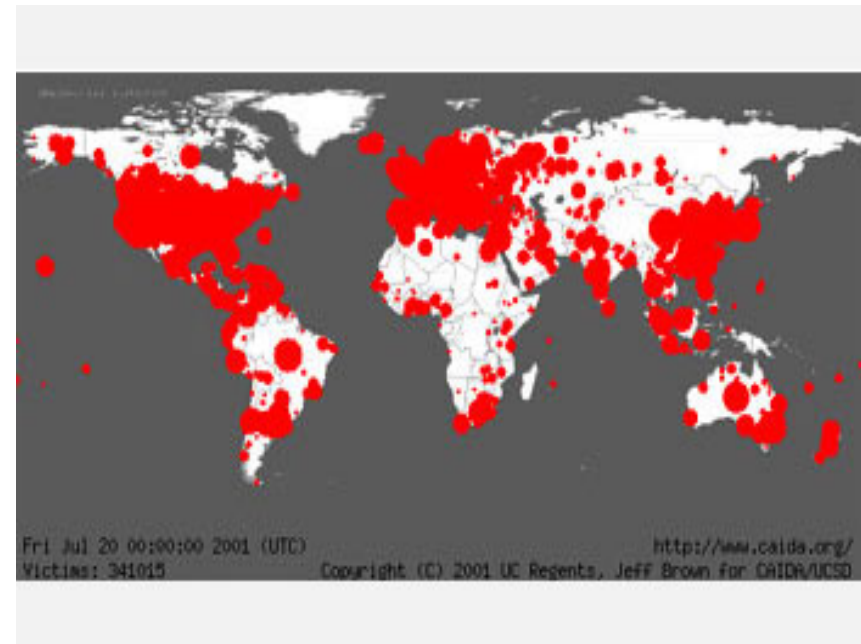
SMARTPHONES ON THE RISE



GROWTH IN MOBILE MALWARE



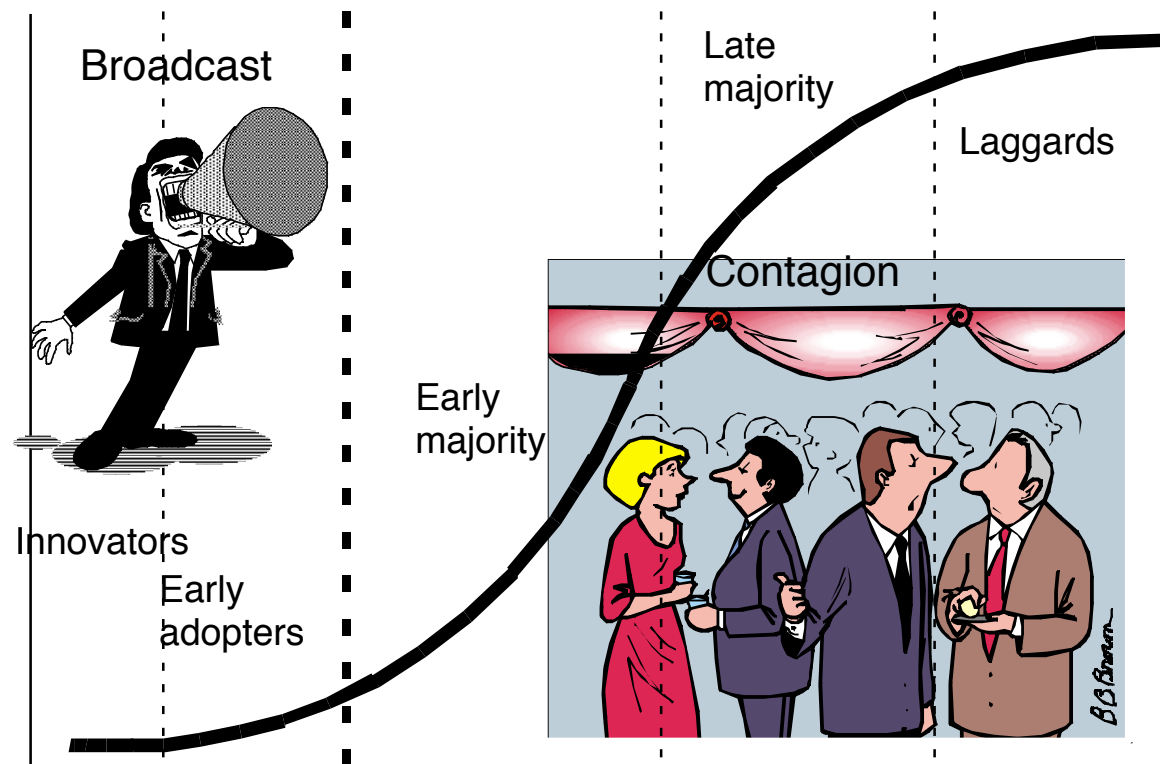
Code Red Worm paralyzed many countries' Internet



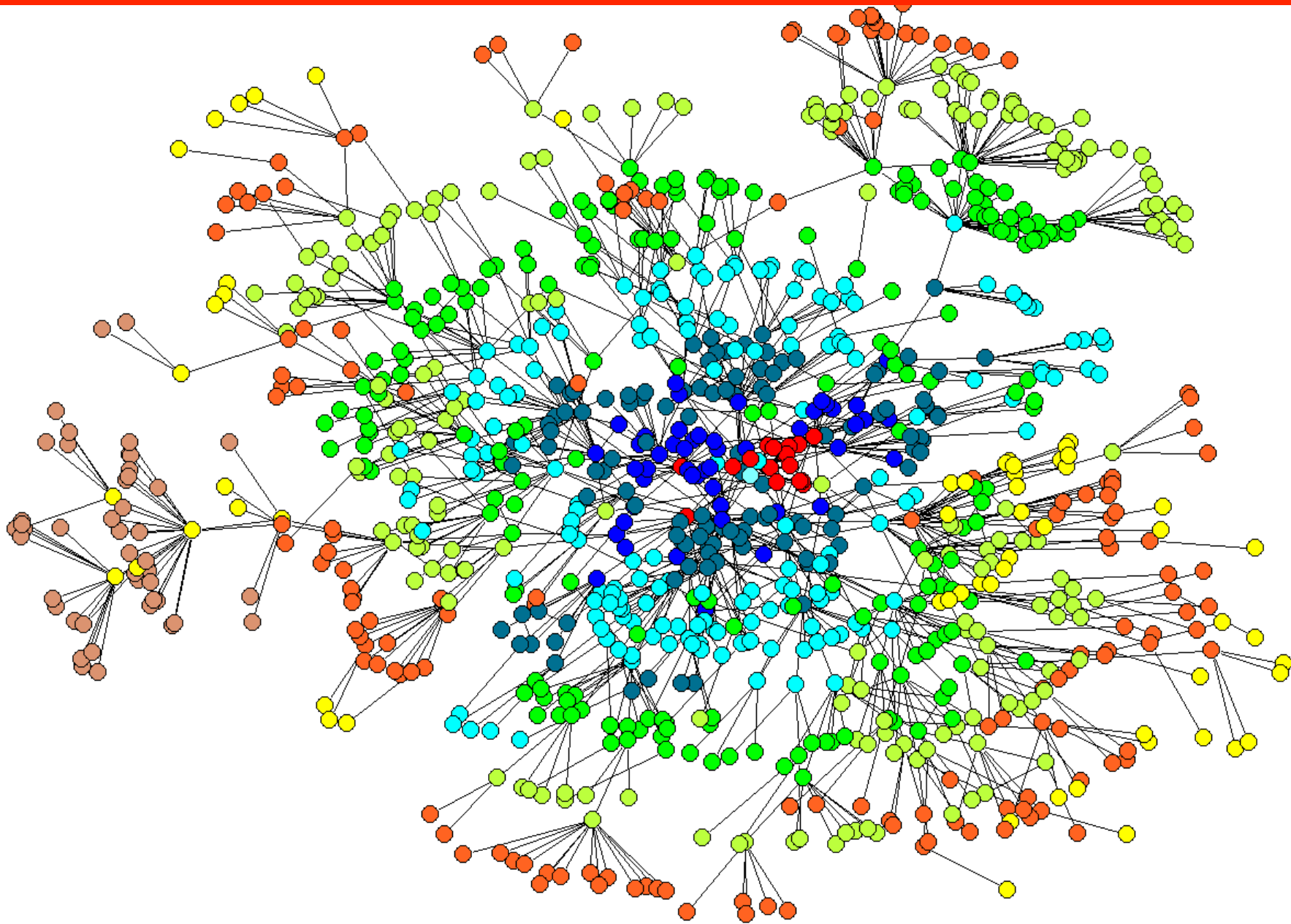
<http://www.caida.org/publications/visualizations/>

Hypponen M. *Scientific American* Nov. 70-77 (2006).

Diffusion of Innovation – The Adoption Curve



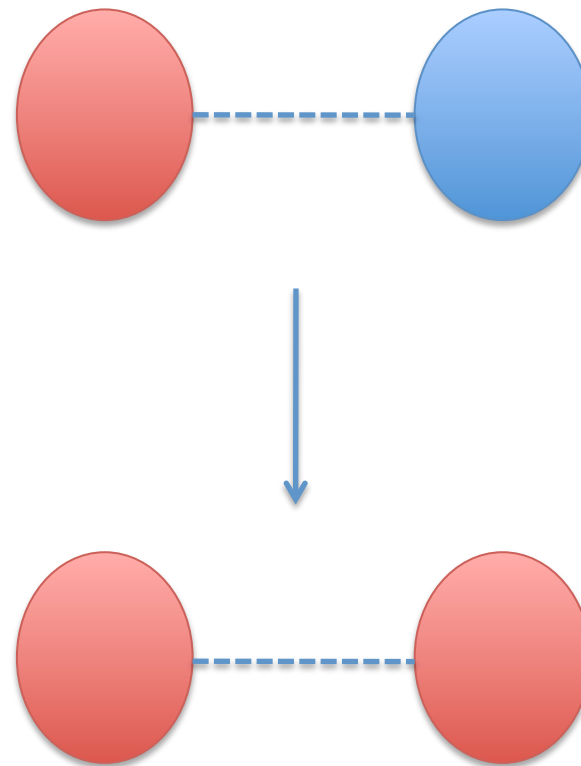
Information Spreading



Epidemic Spreading – Network

- Epidemic spreading always implies network structure!

Spreading happens only when the carries of the diseases/virus/idea are **connected to each other**.



Epidemic Spreading – Network

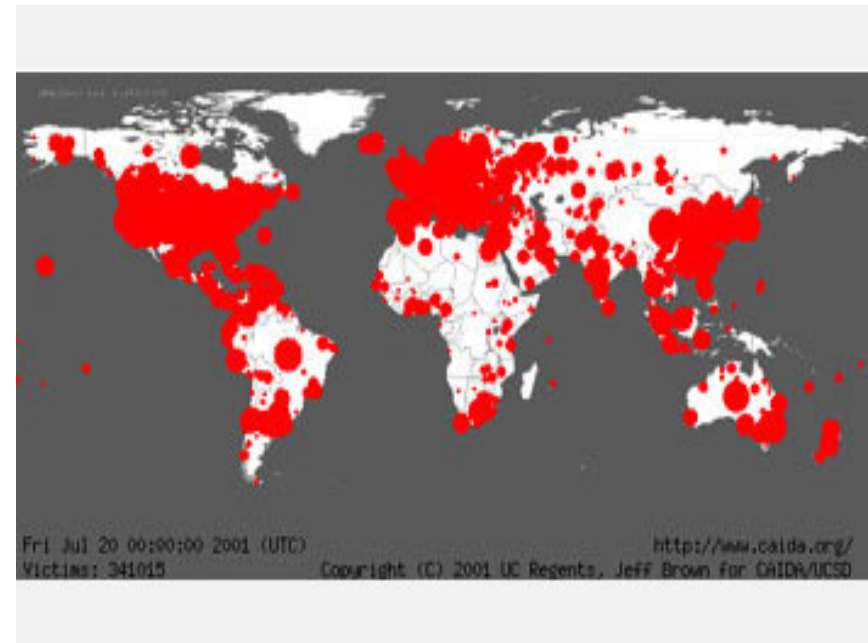


Epidemic Spreading – Network



The transportation network

L. Hufnagel et al. *PNAS* **101**, 15124 (2004)



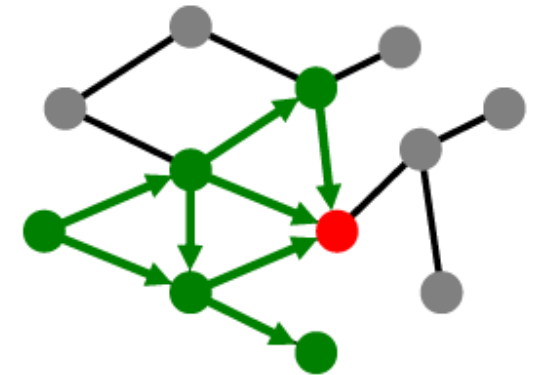
Internet

<http://www.caida.org/publications/visualizations/>

How to model diffusion?

■ Probabilistic models:

- Models of influence or disease spreading
 - An infected node tries to “push” the contagion to an uninfected node
- **Example:**
 - You “catch” a disease with some prob. from each active neighbor in the network



■ Decision based models:

- Models of product adoption, decision making
 - A node observes decisions of its neighbors and makes its own decision
- **Example:**
 - You join demonstrations if k of your friends do so too

Decision-based diffusion models

Collective action
[Granovetter 1978]

- **Collective Action** [Granovetter, '78]
 - Model where everyone sees everyone else's behavior
 - **Examples:**
 - Clapping or getting up and leaving in a theater
 - Keeping your money or not in a stock market
 - Neighborhoods in cities changing ethnic composition
 - Riots, protests, strikes

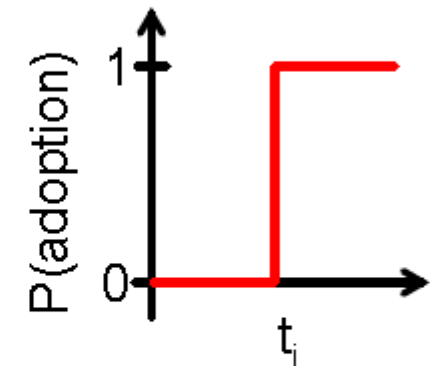
The model of collective action

- **n people – everyone observes all actions**

- Each person i has a threshold t_i

- Node i will adopt the behavior iff at least t_i other people are adopters:

- Small t_i : early adopter
- Large t_i : late adopter



- **The population is described by $\{t_1, \dots, t_n\}$**

- **$F(x)$... fraction of people with threshold $t_i \leq x$**

Dynamics of collective action

- **Think of the step-by-step change in number of people adopting the behavior:**

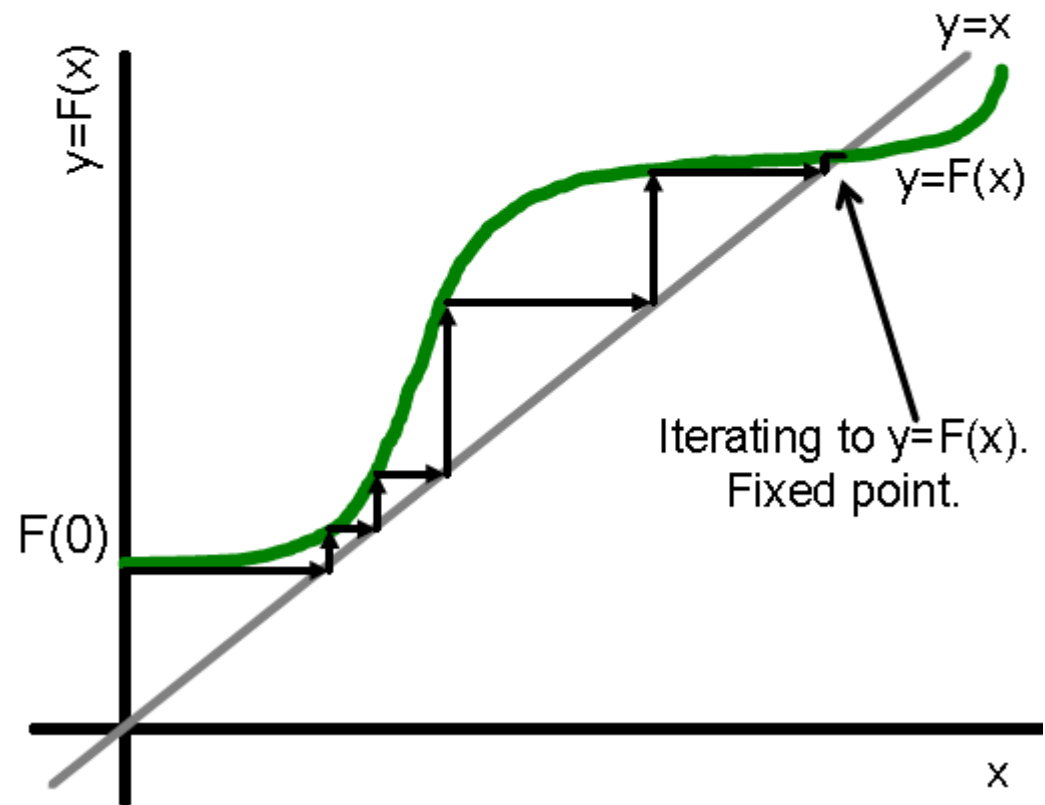
- $F(x)$... fraction of people with threshold $\leq x$
- $s(t)$... number of participants at time t

- **Easy to simulate:**

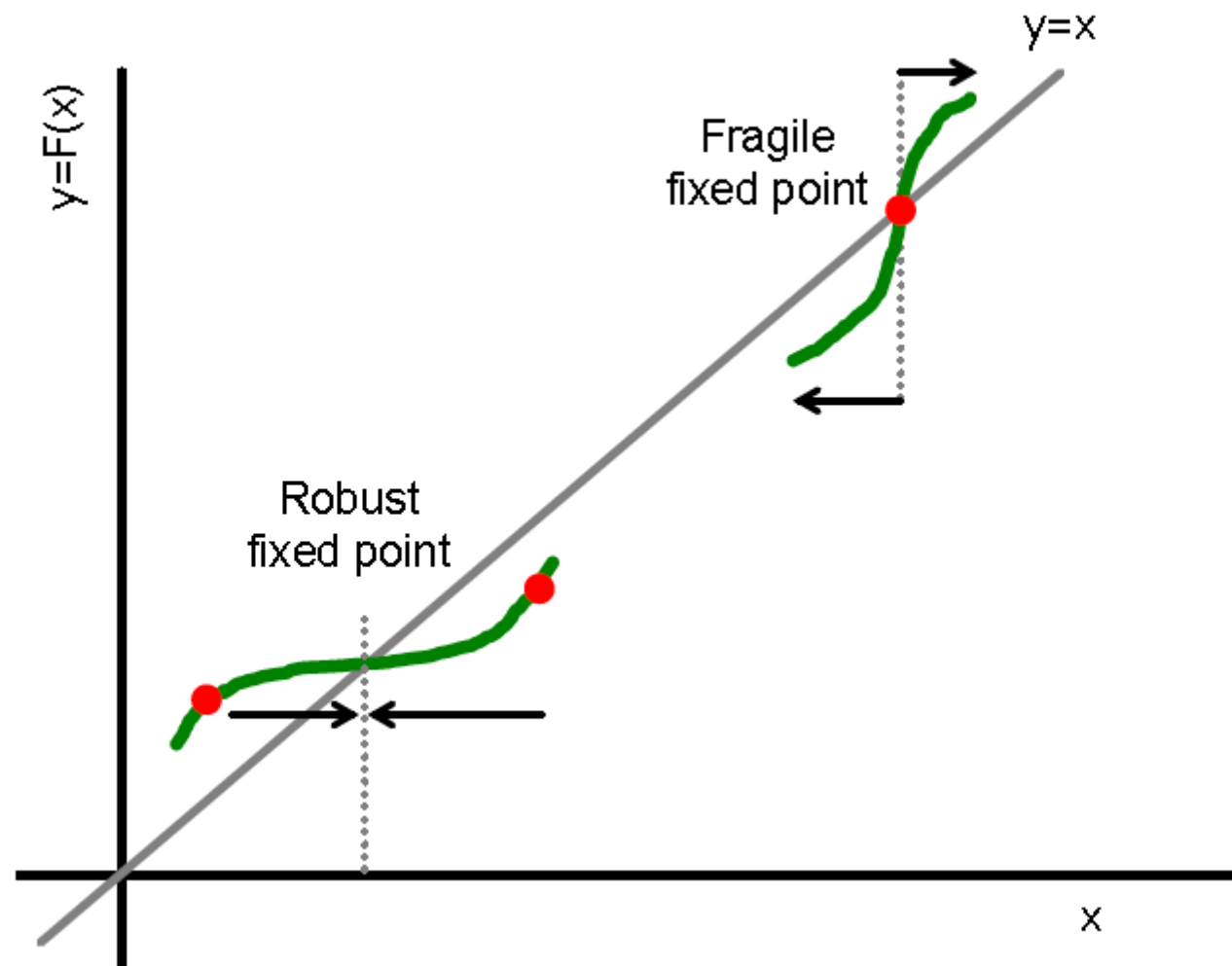
- $s(0) = 0$
- $s(1) = F(0)$
- $s(2) = F(s(1)) = F(F(0))$
- $s(t+1) = F(s(t)) = F^{t+1}(0)$

- **Fixed point: $F(x)=x$**

- There could be other fixed points but starting from 0 we never reach them



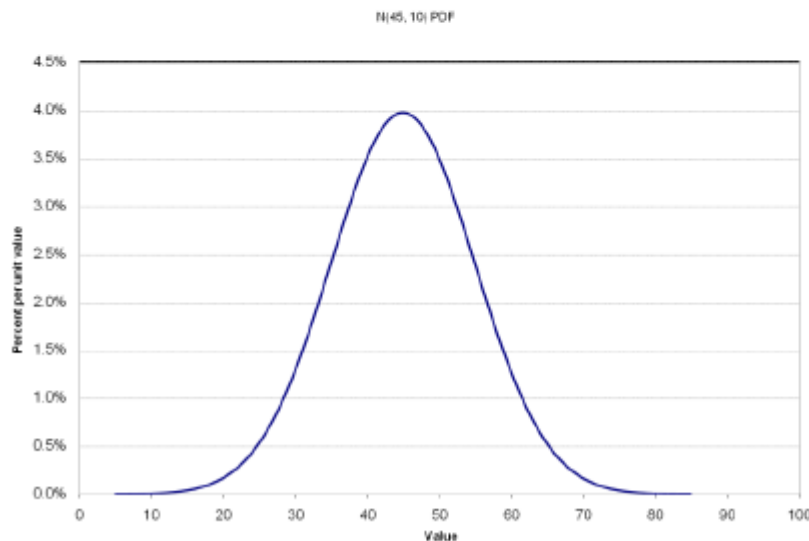
Fragile vs. robust fixed points



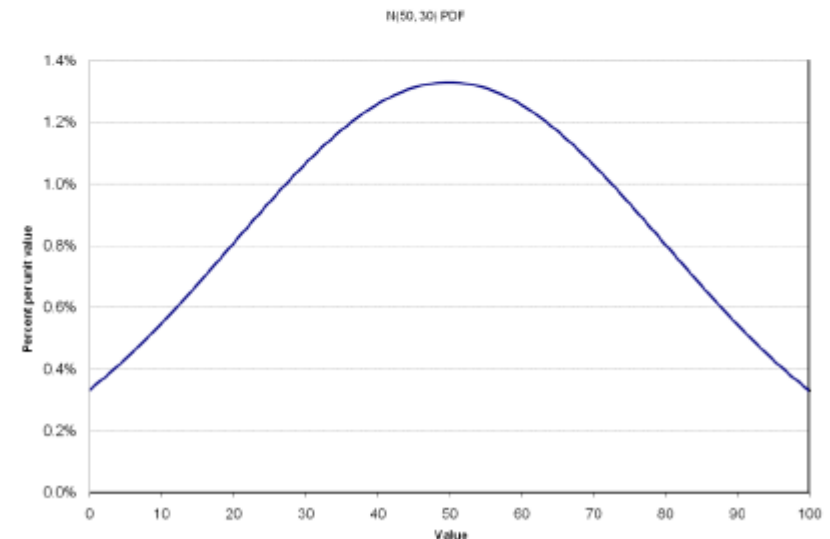
Distribution of thresholds (trust)

- Each threshold t_i is drawn independently from some distribution $F(x) = \Pr[\text{thresh} \leq x]$
 - **Suppose:** Normal with $\mu=n/2$, variance σ

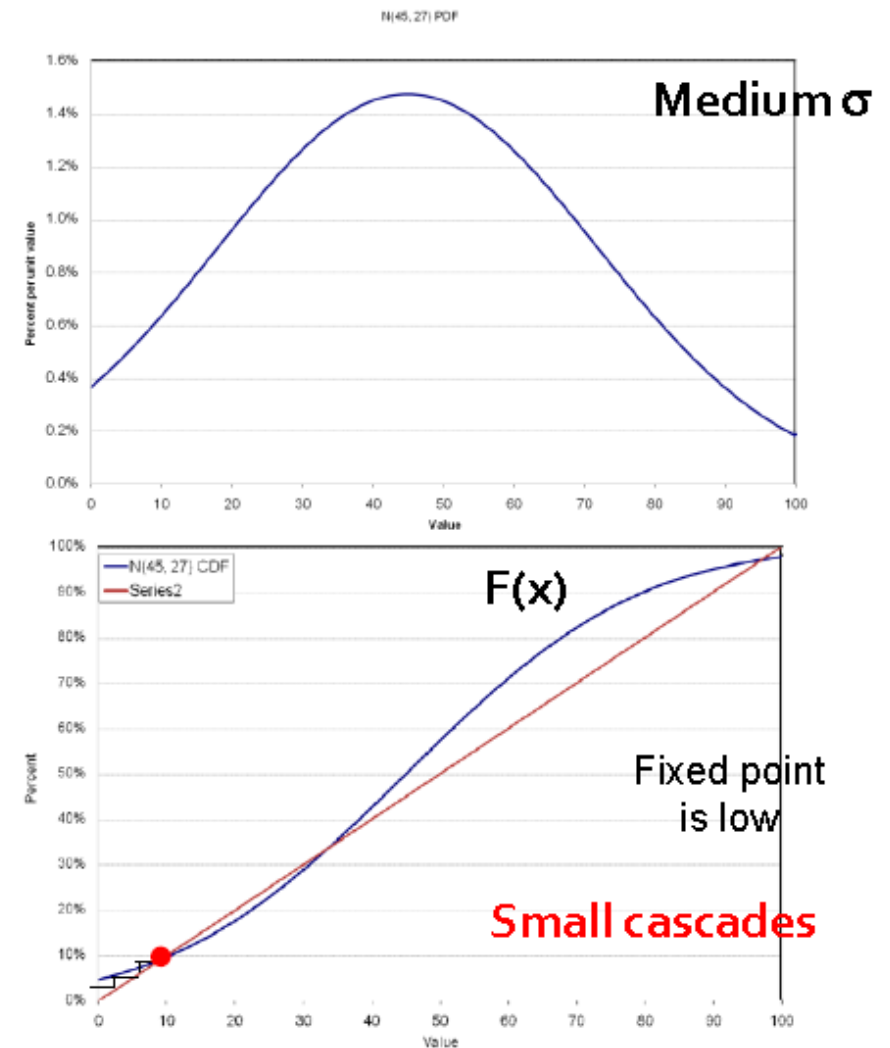
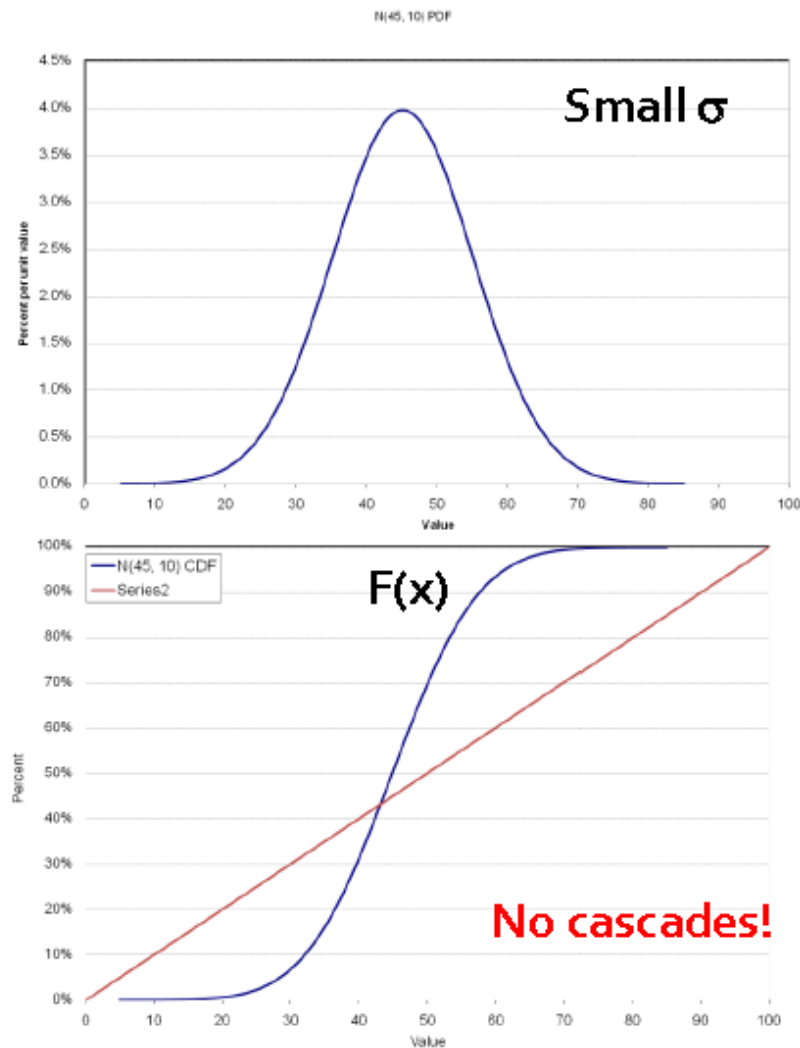
Small σ :



Large σ :

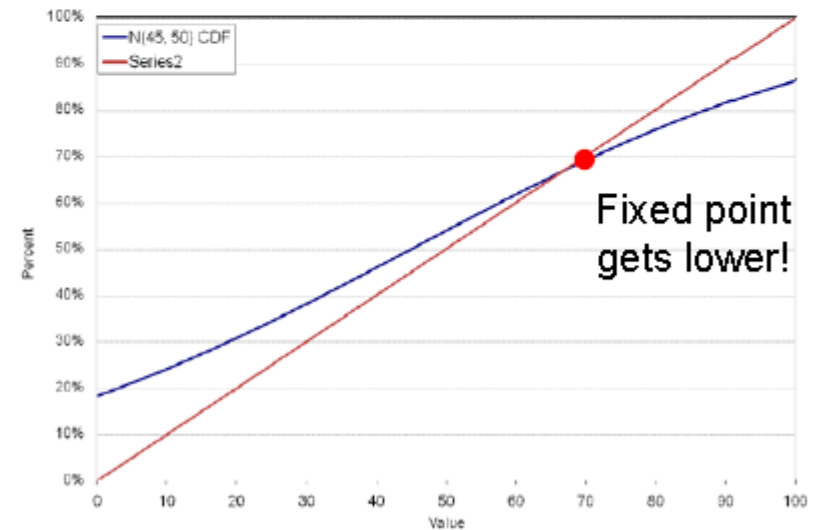
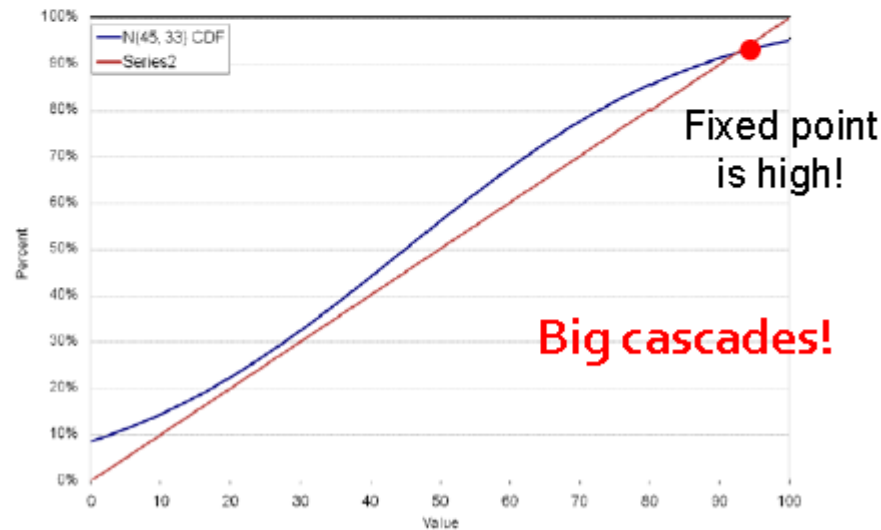
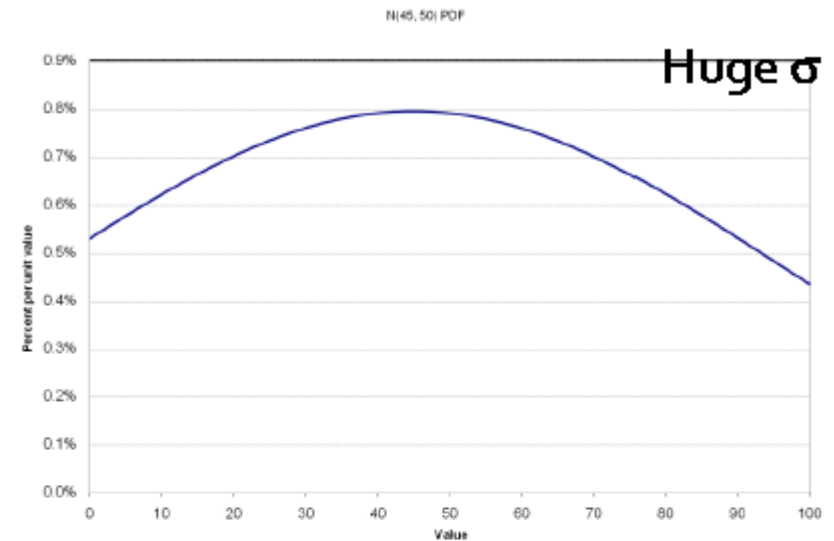
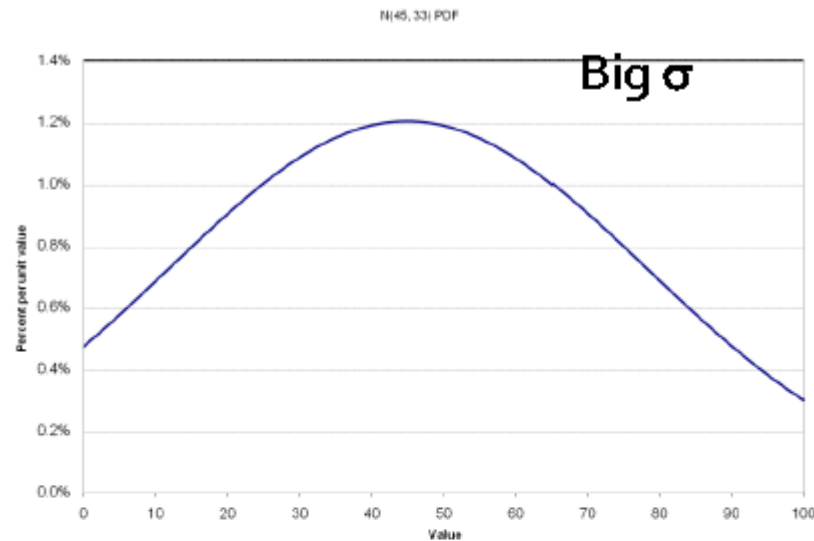


Simulation

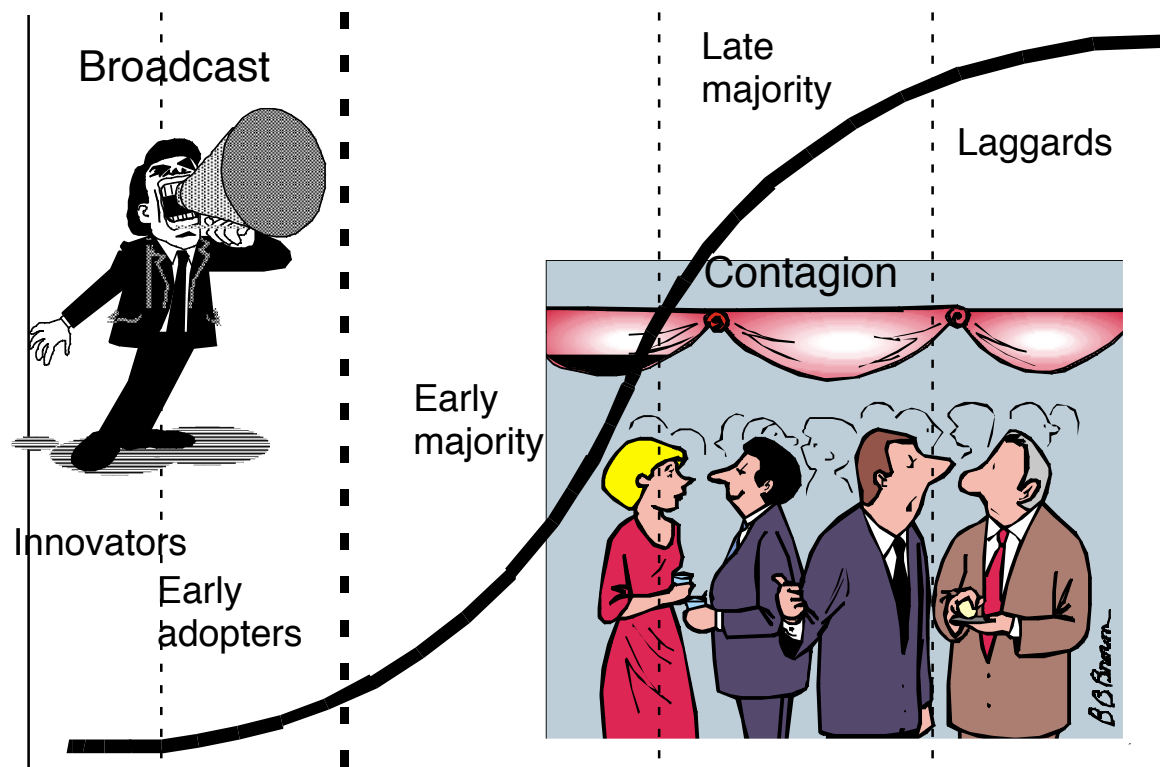


Bigger variance let's you build a bridge from early adopters to mainstream

Simulation



But if we increase the variance even more we move the higher fixed point lower



Weaknesses of the CA model

- **It does not take into account:**
 - No notion of social network – more influential users
 - It matters who the early adopters are, not just how many
 - Models people's awareness of size of participation not just actual number of people participating
 - **Modeling thresholds**
 - Richer distributions
 - Deriving thresholds from more basic assumptions
 - game theoretic models

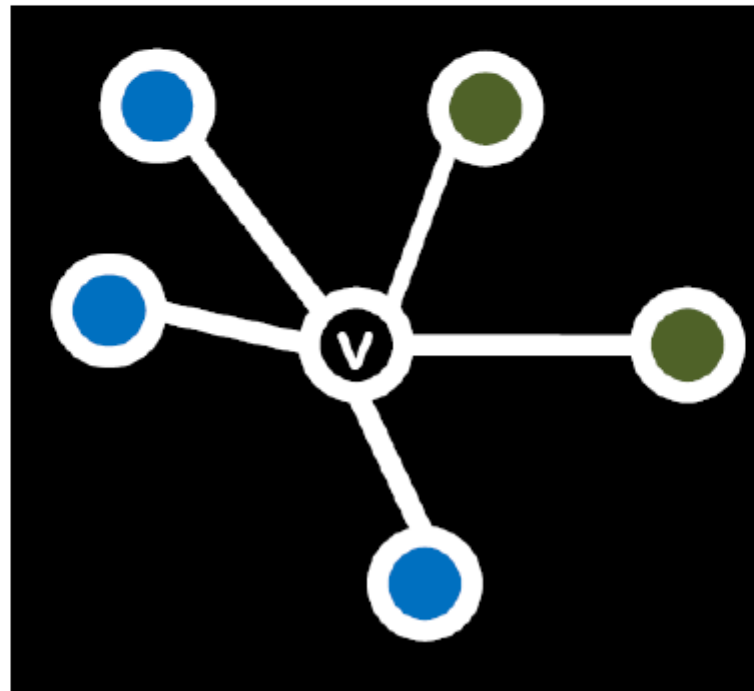
Decision-based diffusion models

Game-theoretic models of cascades

[Moore 2000]

Game theoretic models of cascades

- Based on 2 player coordination game
 - 2 players – each chooses technology A or B
 - Each person can only adopt **one** “behavior”, A or B
 - You gain more payoff if your friend has adopted the **same** behavior as you



Local view of the network of node v

Rules of the game

- **Payoff matrix:**

- If both v and w adopt behavior A , they each get payoff $a > 0$
- If v and w adopt behavior B , they each get payoff $b > 0$
- If v and w adopt the opposite behaviors, they each get 0

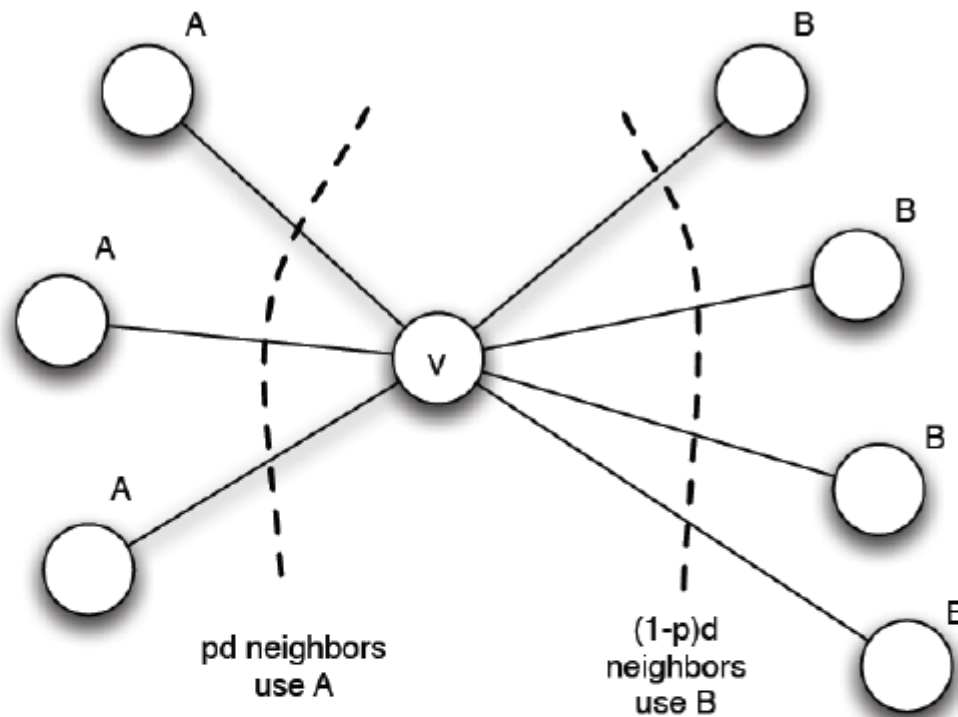


- **In some large network:**

- Each node v is playing a copy of the game with each of its neighbors
- **Payoff:** sum of node payoffs per game

		w	
		A	B
v	A	a, a	$0, 0$
	B	$0, 0$	b, b

Decision rule for node v



Threshold:
 v chooses A if $p > q$

$$q = \frac{b}{a + b}$$

- Let v have d neighbors
- Assume fraction p of v 's neighbors adopt A
 - $\text{Payoff}_v = a \cdot p \cdot d$ if v chooses A
 - $= b \cdot (1-p) \cdot d$ if v chooses B
- **Thus: v chooses A if: $a \cdot p \cdot d > b \cdot (1-p) \cdot d$**

Example

- **Scenario:**

Graph where everyone starts with B.

Small set S of early adopters of A

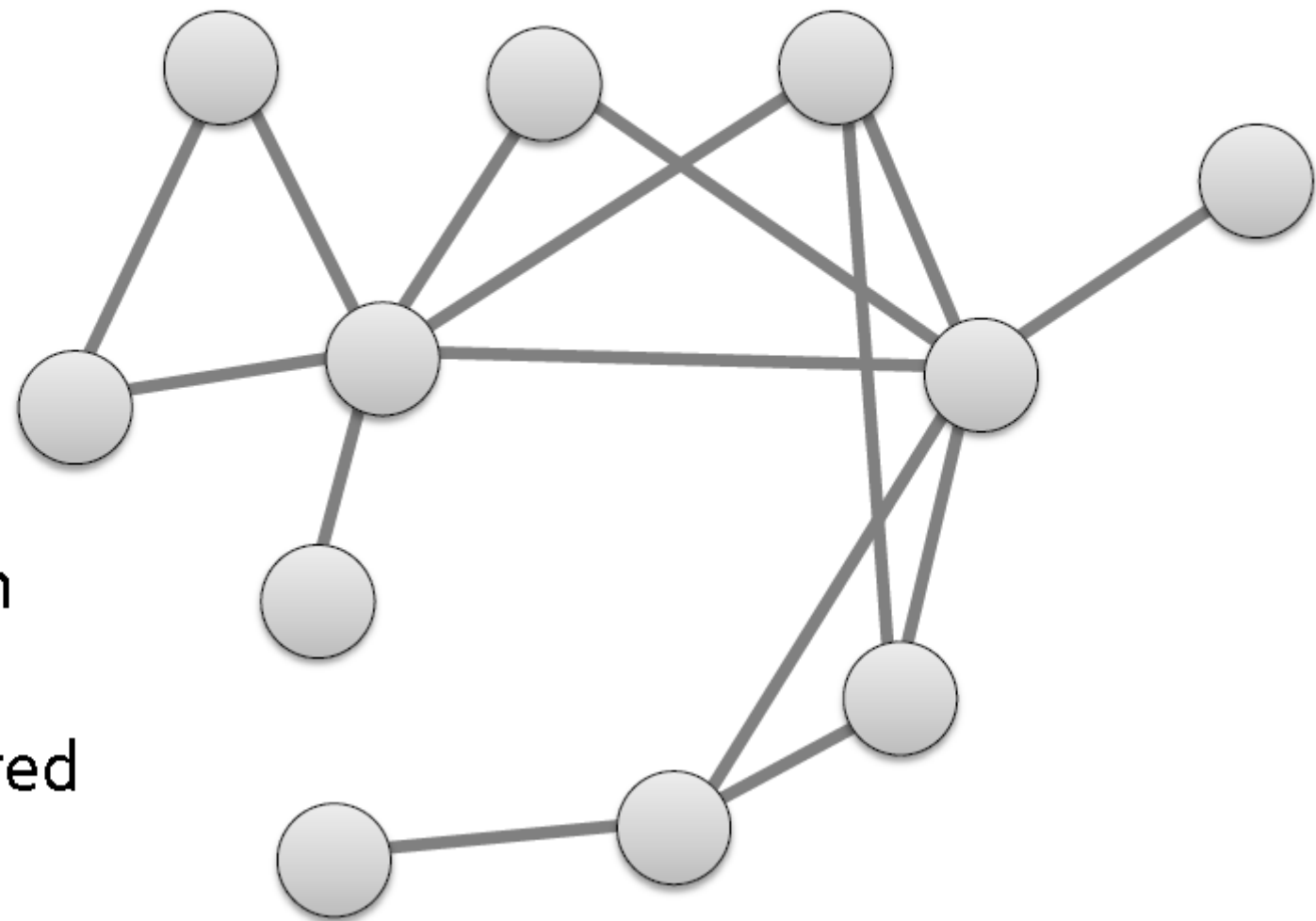
- Hard wire S – they keep using A no matter what payoffs tell them to do

- Payoffs are set in such a way that nodes say:

If at least 50% of my friends are red I'll be red

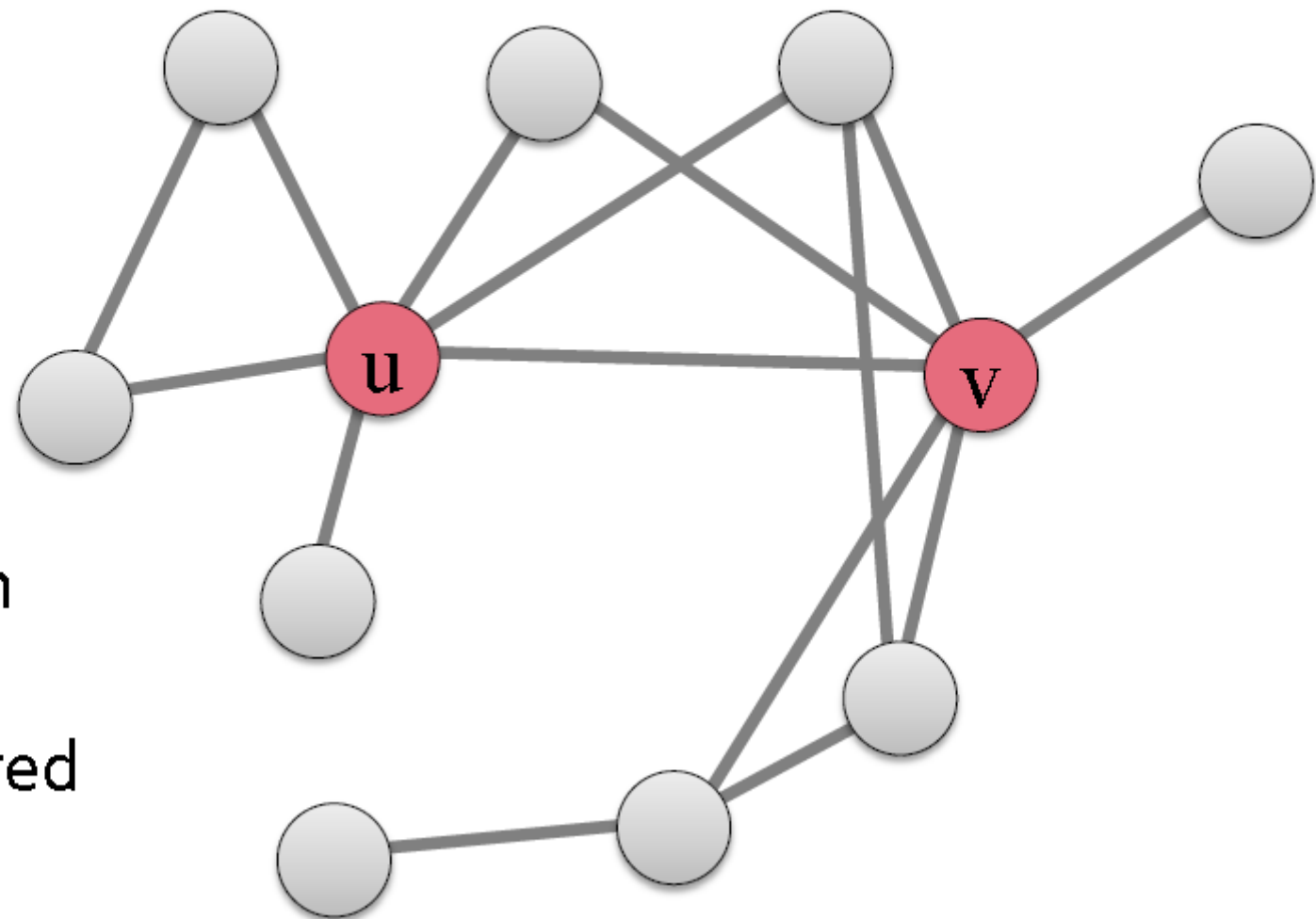
(this means: $a = b + \epsilon$)

$$S = \{u, v\}$$



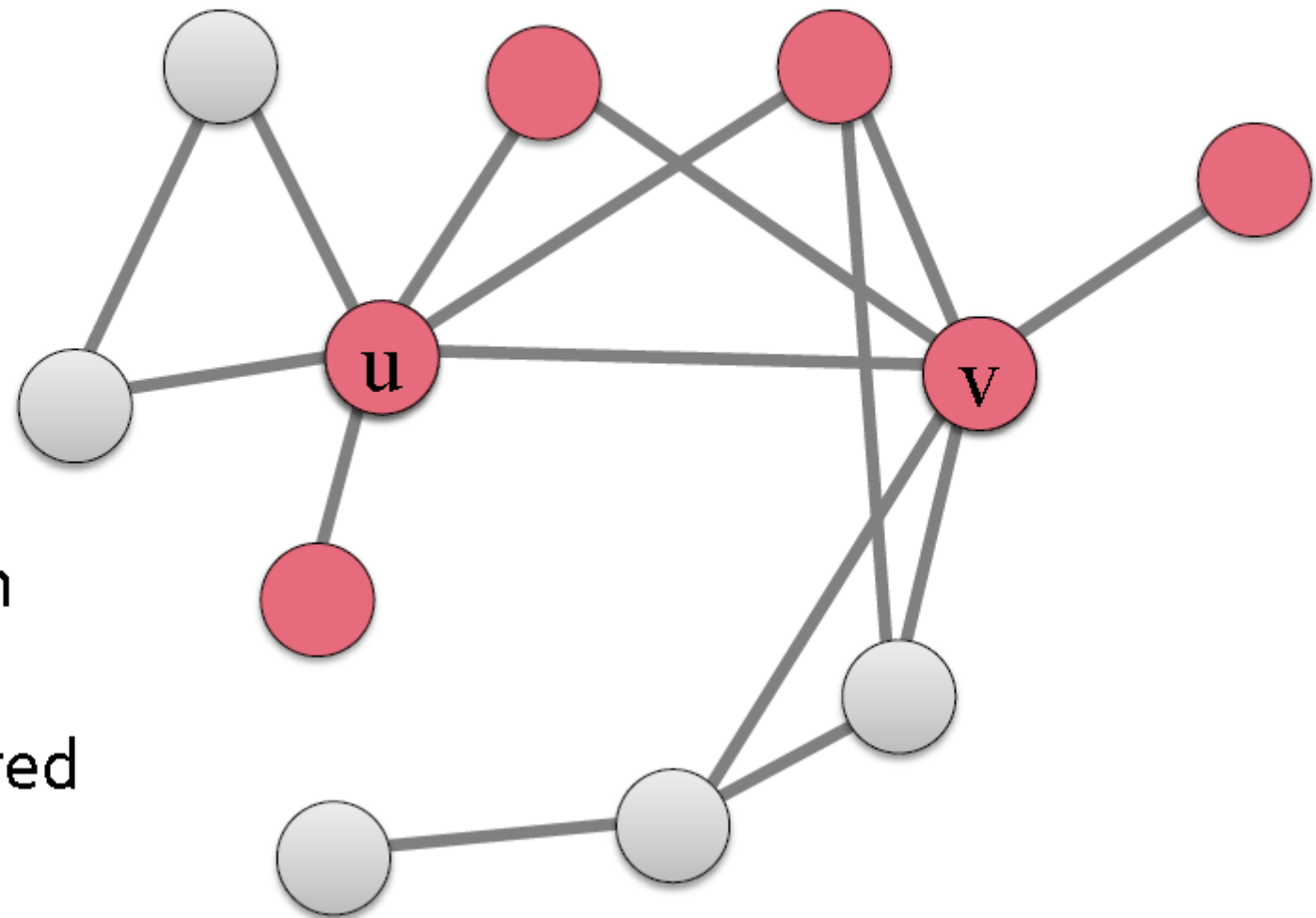
If **more** than
50% of my
friends are red
I'll be red

$$S = \{u, v\}$$



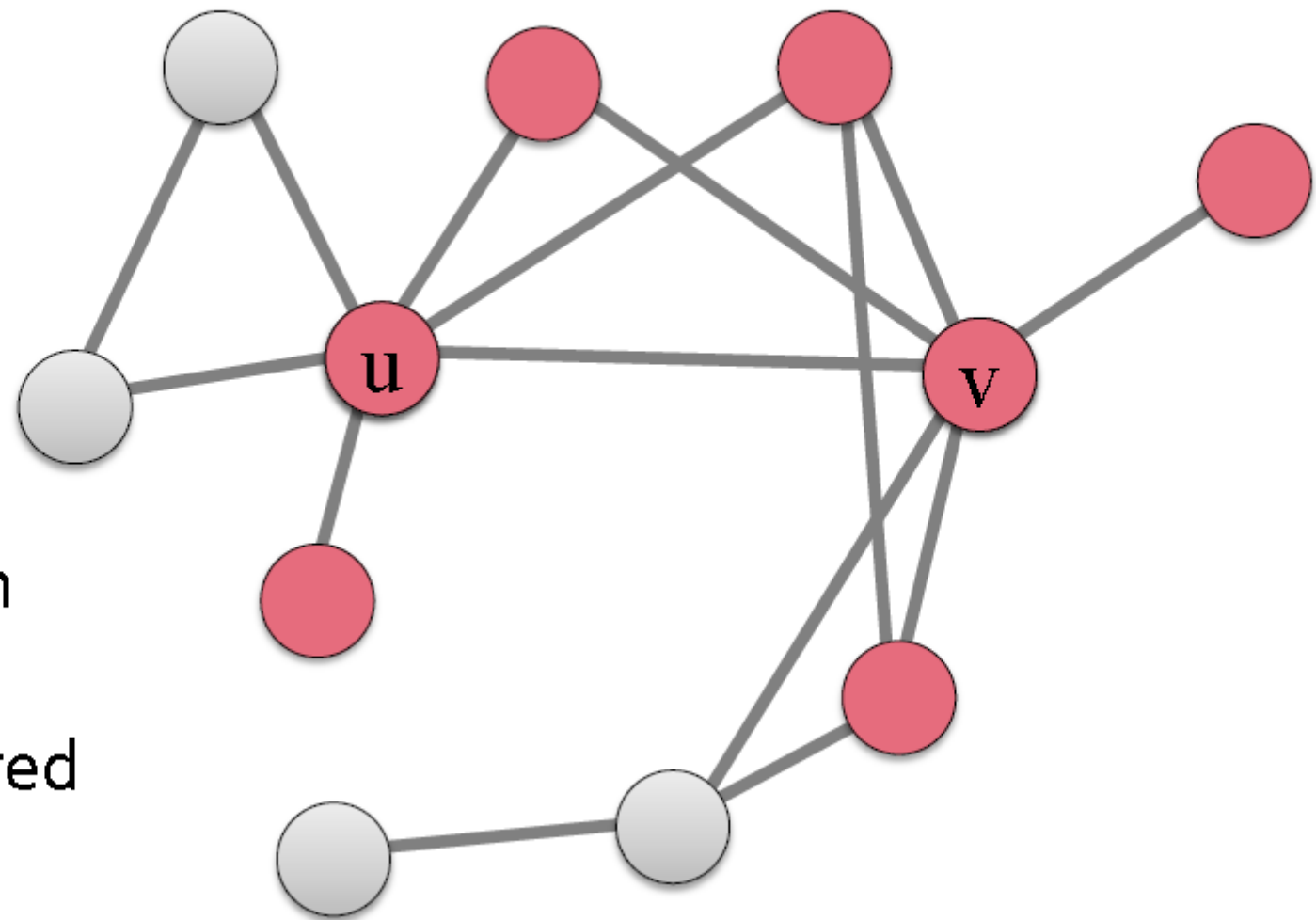
If **more** than
50% of my
friends are red
I'll be red

$$S = \{u, v\}$$



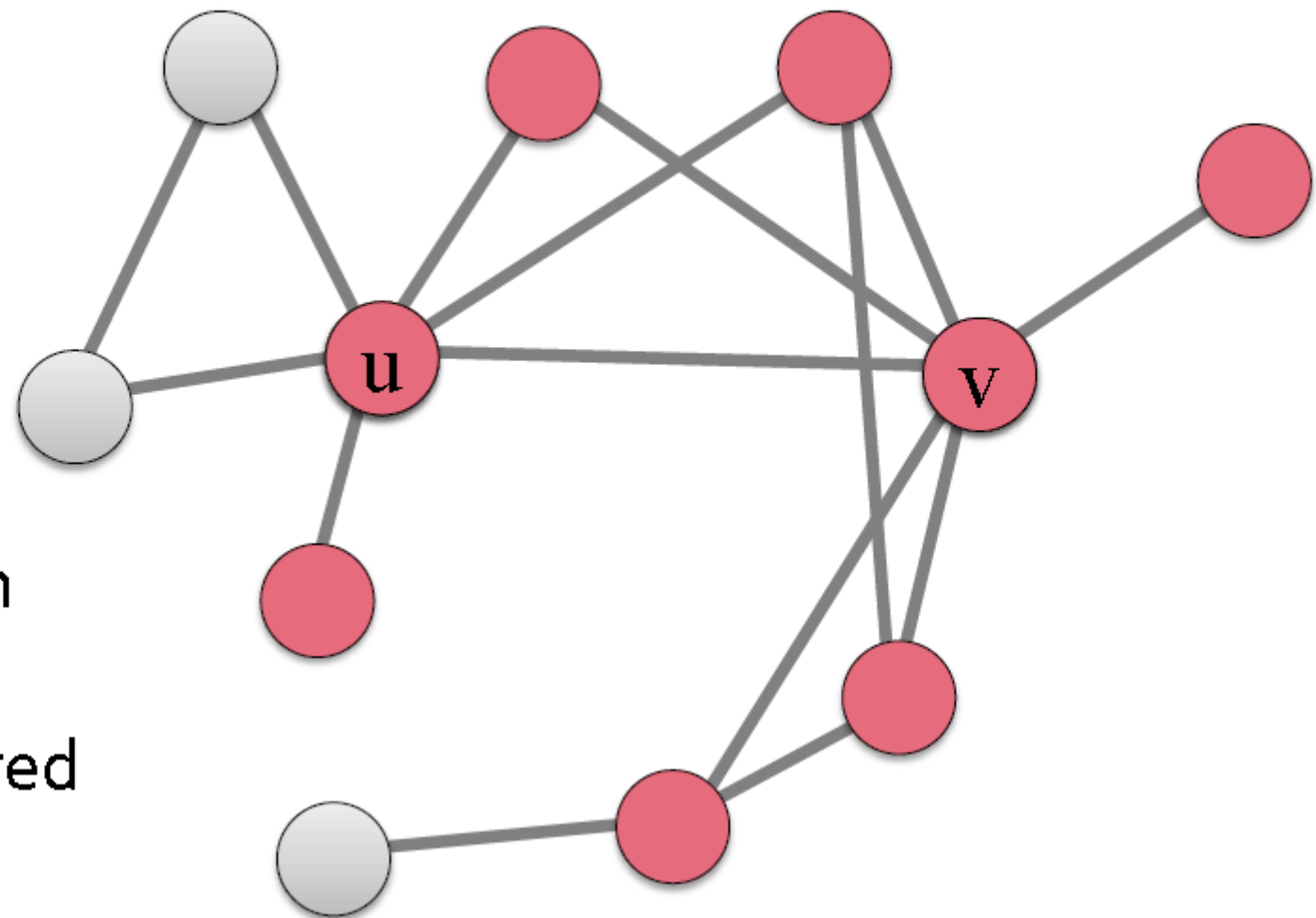
If **more** than
50% of my
friends are red
I'll be red

$$S = \{u, v\}$$



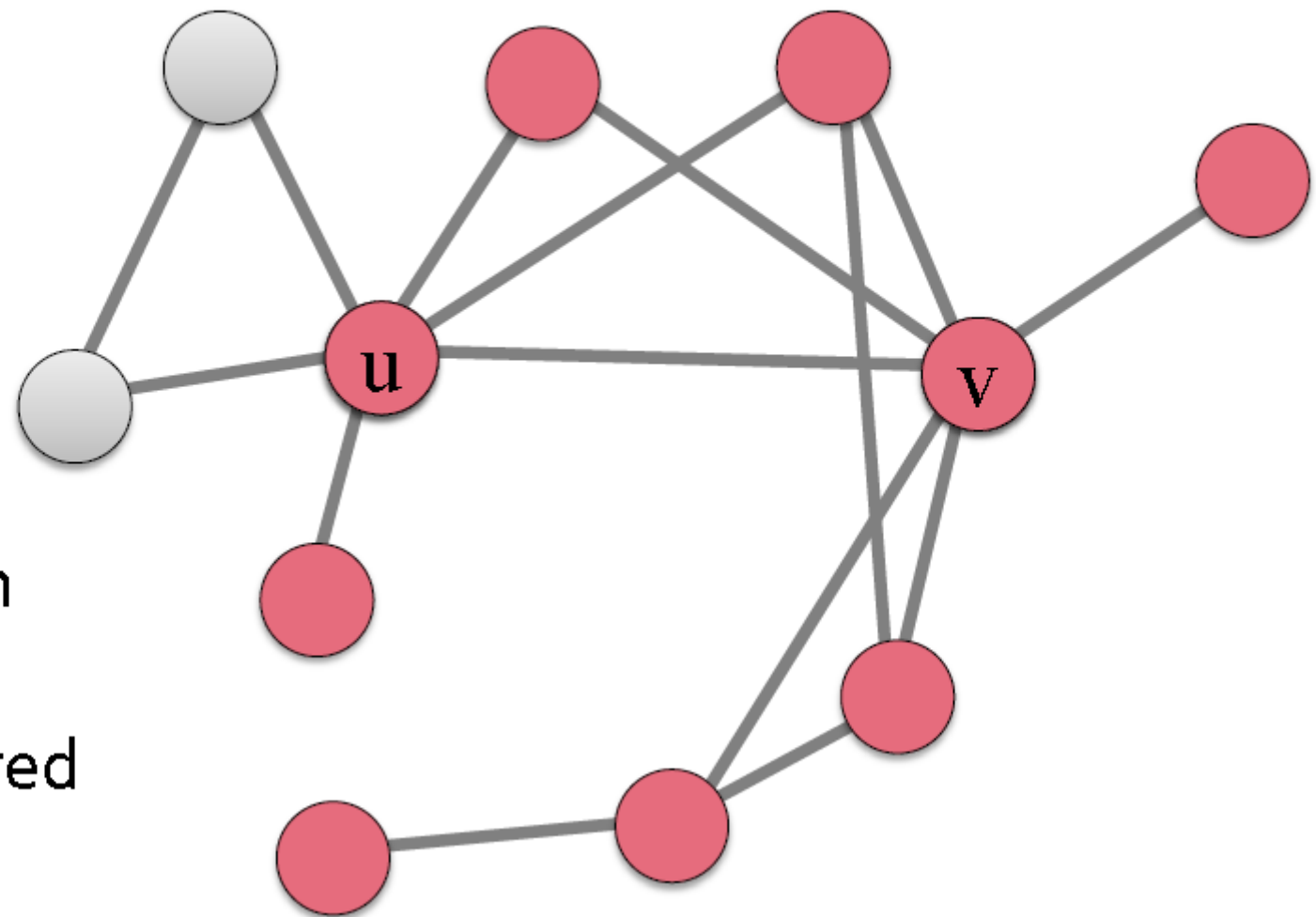
If **more** than
50% of my
friends are red
I'll be red

$$S = \{u, v\}$$



If **more** than
50% of my
friends are red
I'll be red

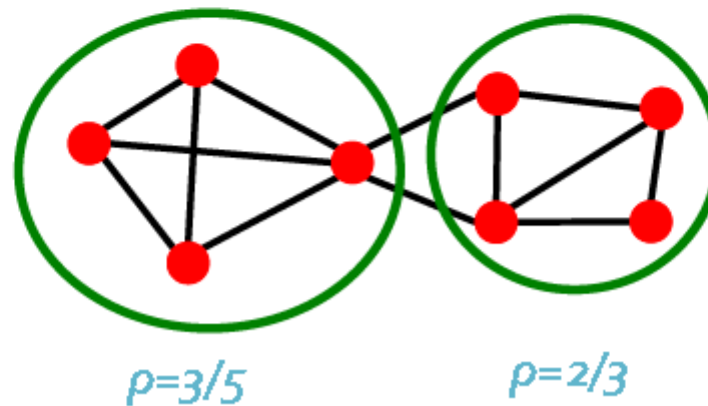
$$S = \{u, v\}$$



If **more** than
50% of my
friends are red
I'll be red

Stopping cascades

- What prevents cascades from spreading?
- Def: **Cluster of density ρ** is a **set of nodes C** where each node in the set has at least ρ fraction of edges in C .



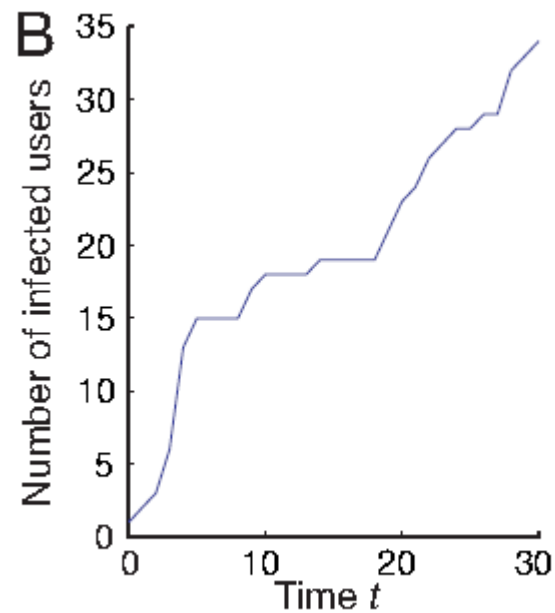
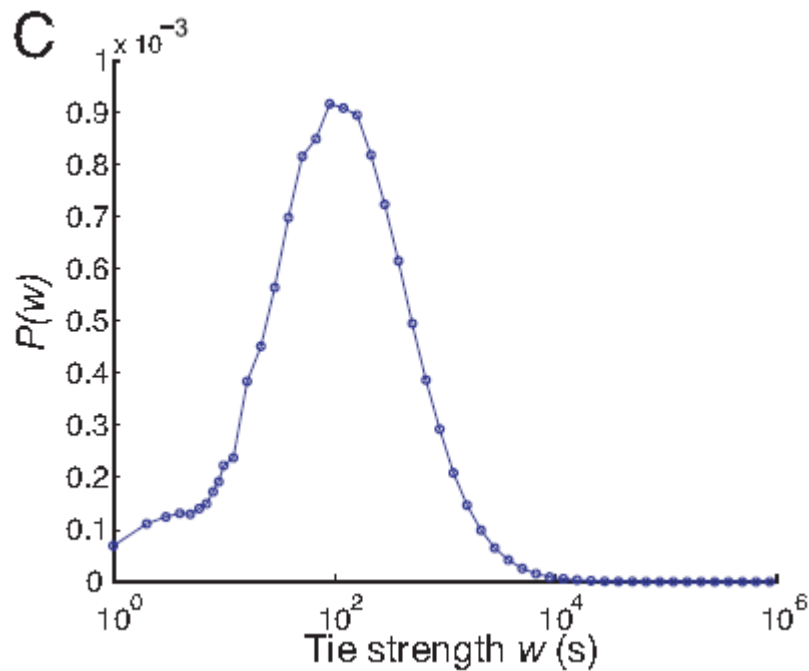
Stopping cascades

- Let S be an initial set of adopters of A
- All nodes apply threshold q to decide whether to switch to A
- **Two facts:**
 - 1) If $G \setminus S$ contains a cluster of density $>(1-q)$ then S can not cause a cascade
 - 2) If S fails to create a cascade, then there is a cluster of density $>(1-q)$ in $G \setminus S$

Empirical studies of cascading behavior

The strength of weak ties ...

- For information **diffusion** (**spreading** of news and rumors on a social network)



The weakness of weak ties

- Diffusion of **innovation / adoption**

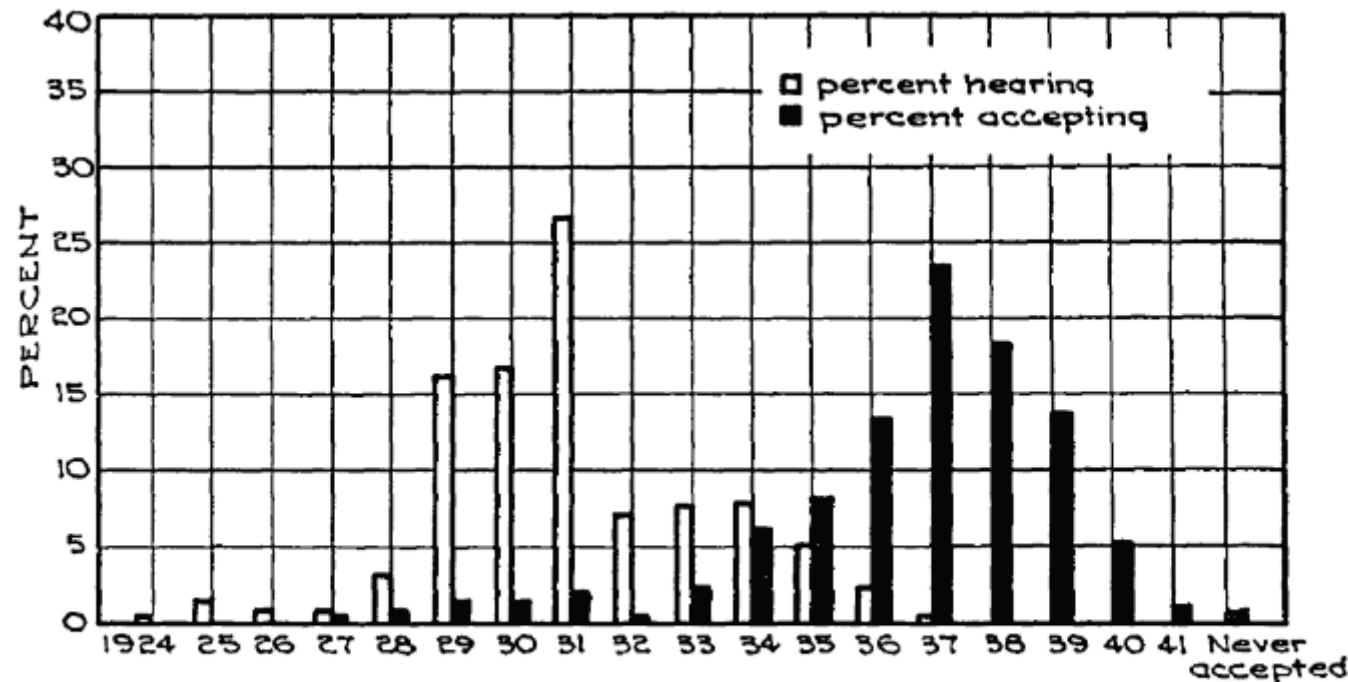
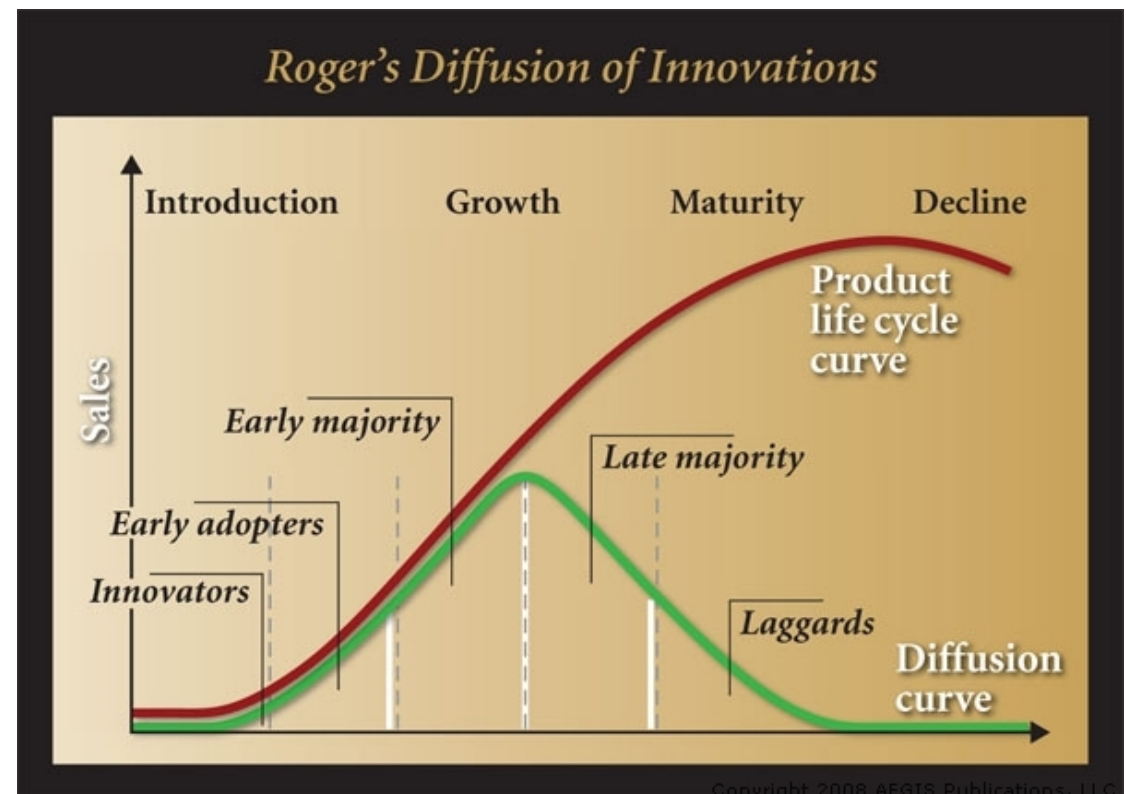
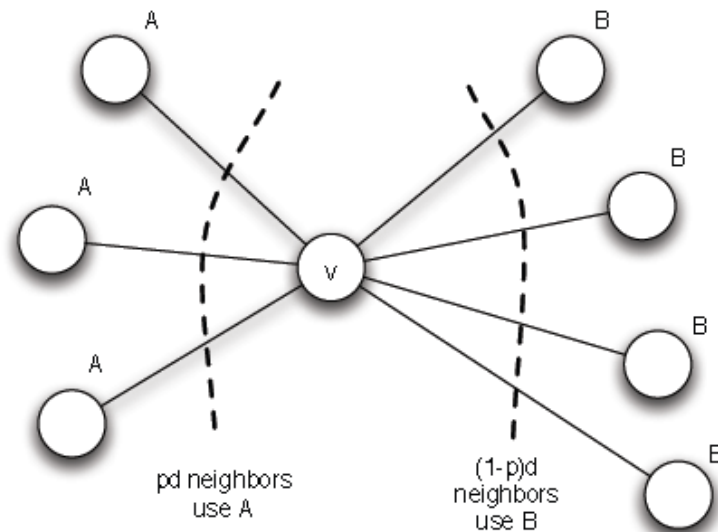


Figure 19.10: The years of first awareness and first adoption for hybrid seed corn in the Ryan-Gross study. (Image from [358].)

The strength of the strong ties for the



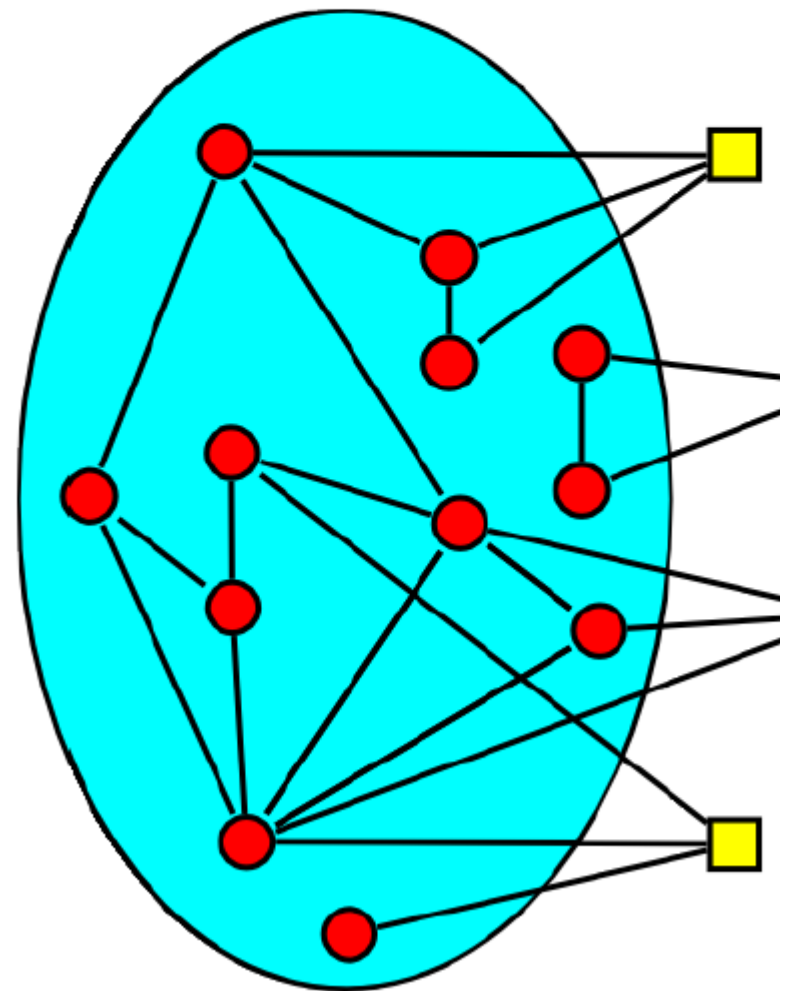
Adoption Curve: LiveJournal

- **Group memberships spread over the network:**

- Red circles represent existing group members
- Yellow squares may join

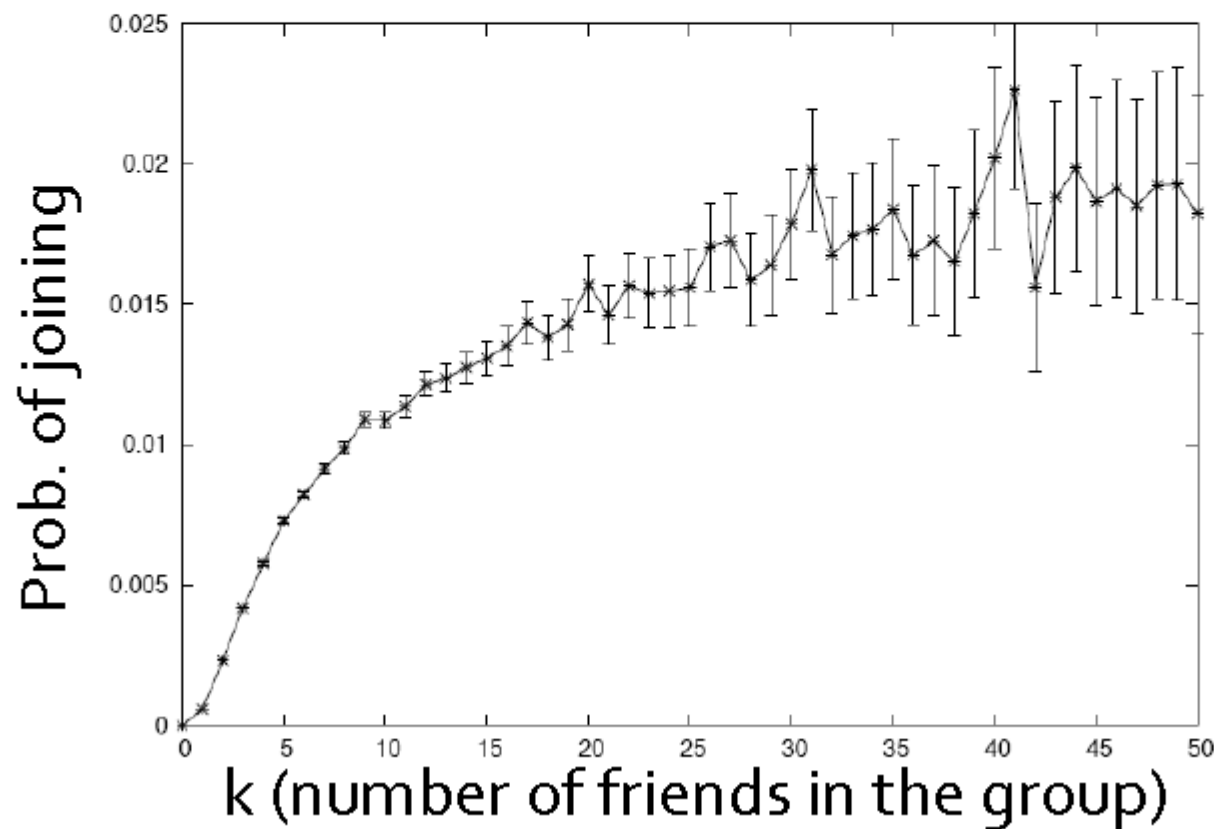
- **Question:**

- How does prob. of joining a group depend on the number of friends already in the group?



Adoption Curve: LiveJournal

- LiveJournal group membership



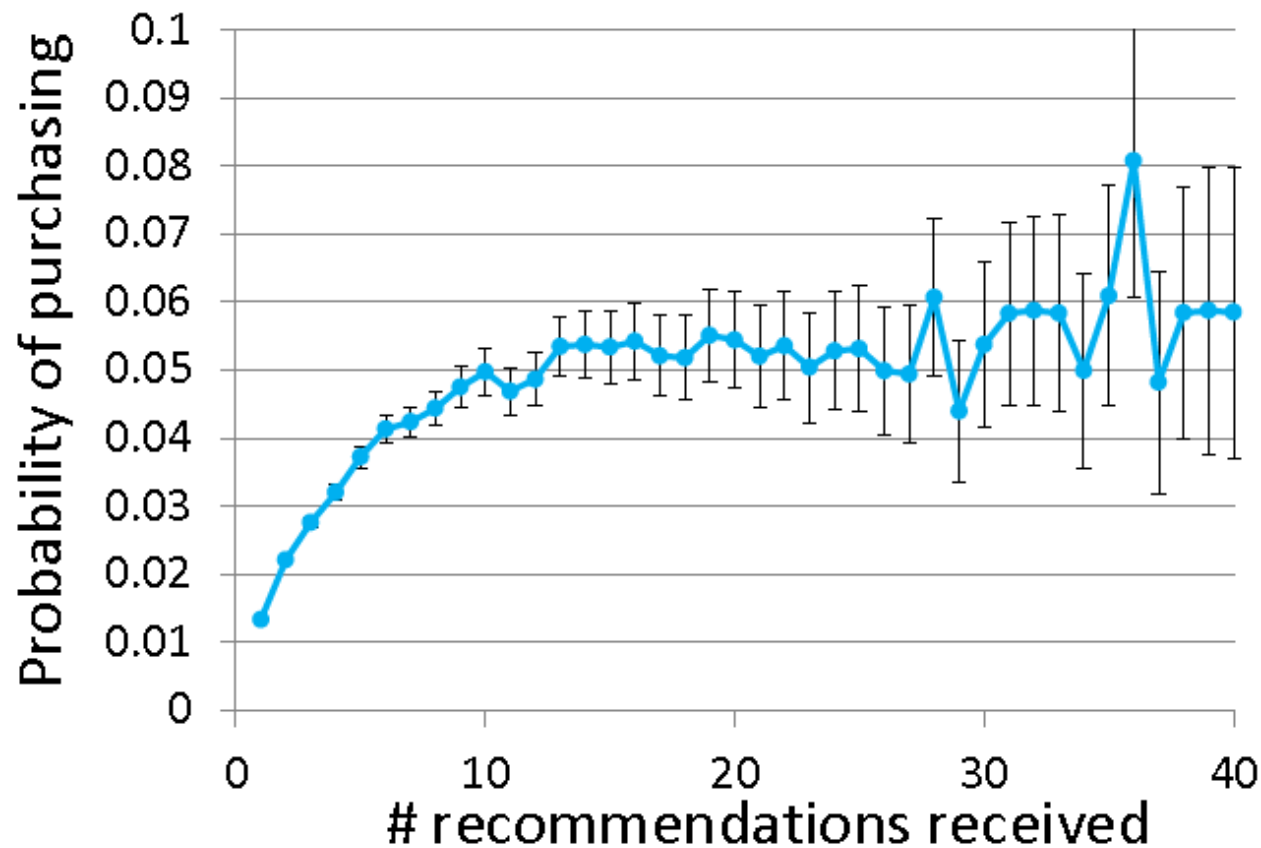
Diffusion in Viral Marketing

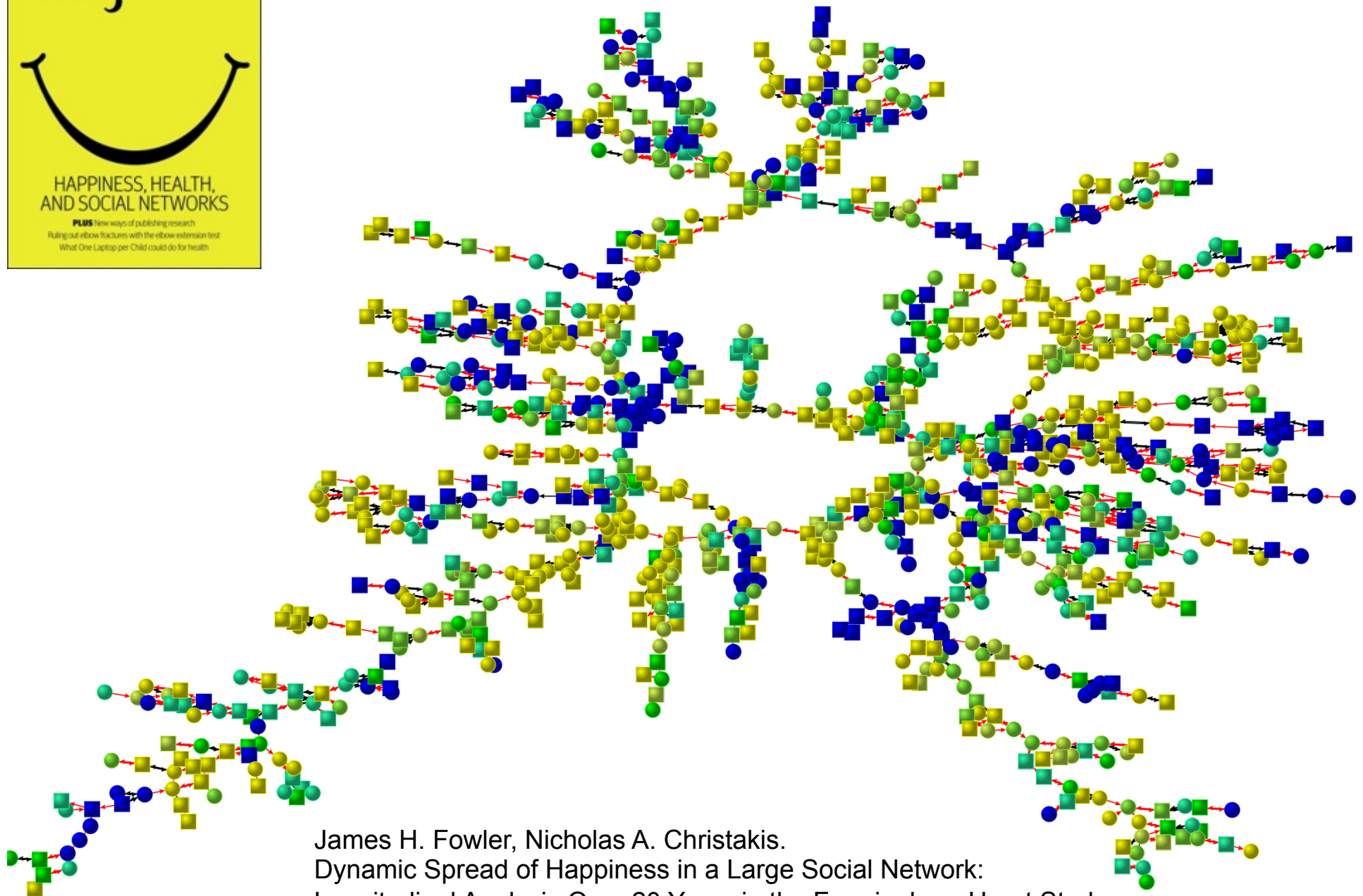
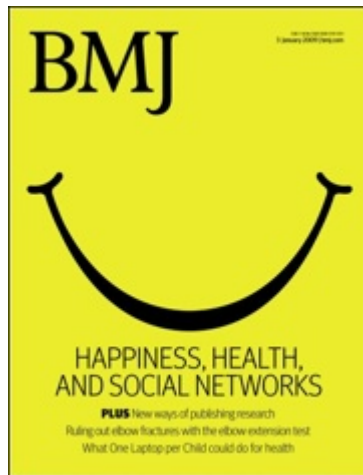
- Senders and followers of recommendations receive discounts on products



- **Data: Incentivized Viral Marketing program**
 - 16 million recommendations
 - 4 million people, 500k products

Adoption Curve: Validation





James H. Fowler, Nicholas A. Christakis.
Dynamic Spread of Happiness in a Large Social Network:
Longitudinal Analysis Over 20 Years in the Framingham Heart Study
British Medical Journal 337 (4 December 2008)

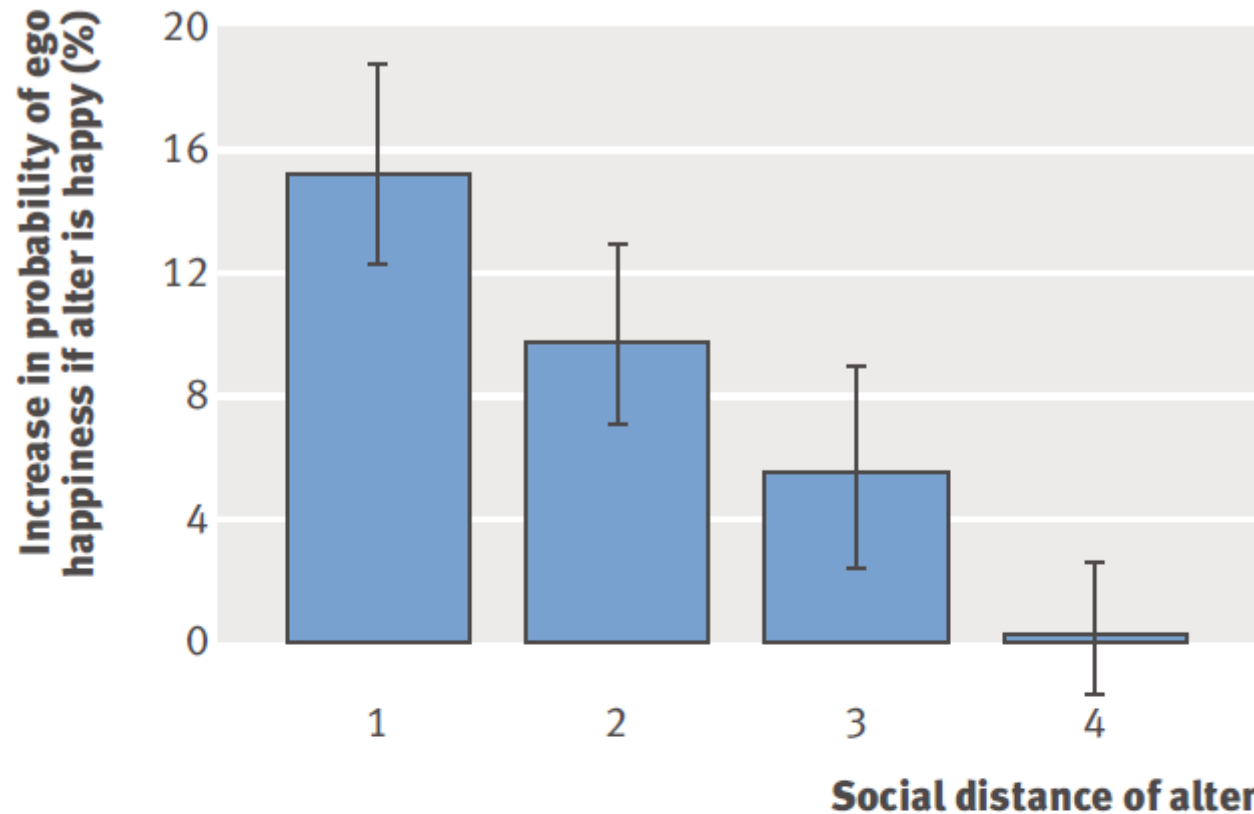
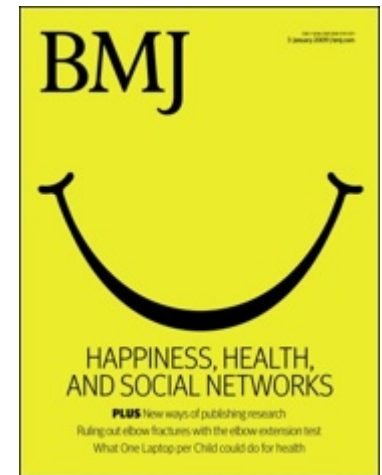


Fig 2 | Social distance and happiness in the Framingham social network. Percentage increase in likelihood an ego is happy if friend or family member at certain social distance is happy (instead of unhappy). The relationship is strongest between individuals who are directly connected but remains significantly >0 at social distances up to three degrees of separation, meaning that a person's happiness is associated with happiness of people up to three degrees removed from them in the network. Values derived by comparing conditional



**Social influence
or
homophily?**



Research highlights

The Three Dimensions of Social Prominence

Diego Pennacchioli^{2,3}, Giulio Rossetti^{1,2}, Luca Pappalardo^{1,2},
Fosca Giannotti², Dino Pedreschi^{1,2}

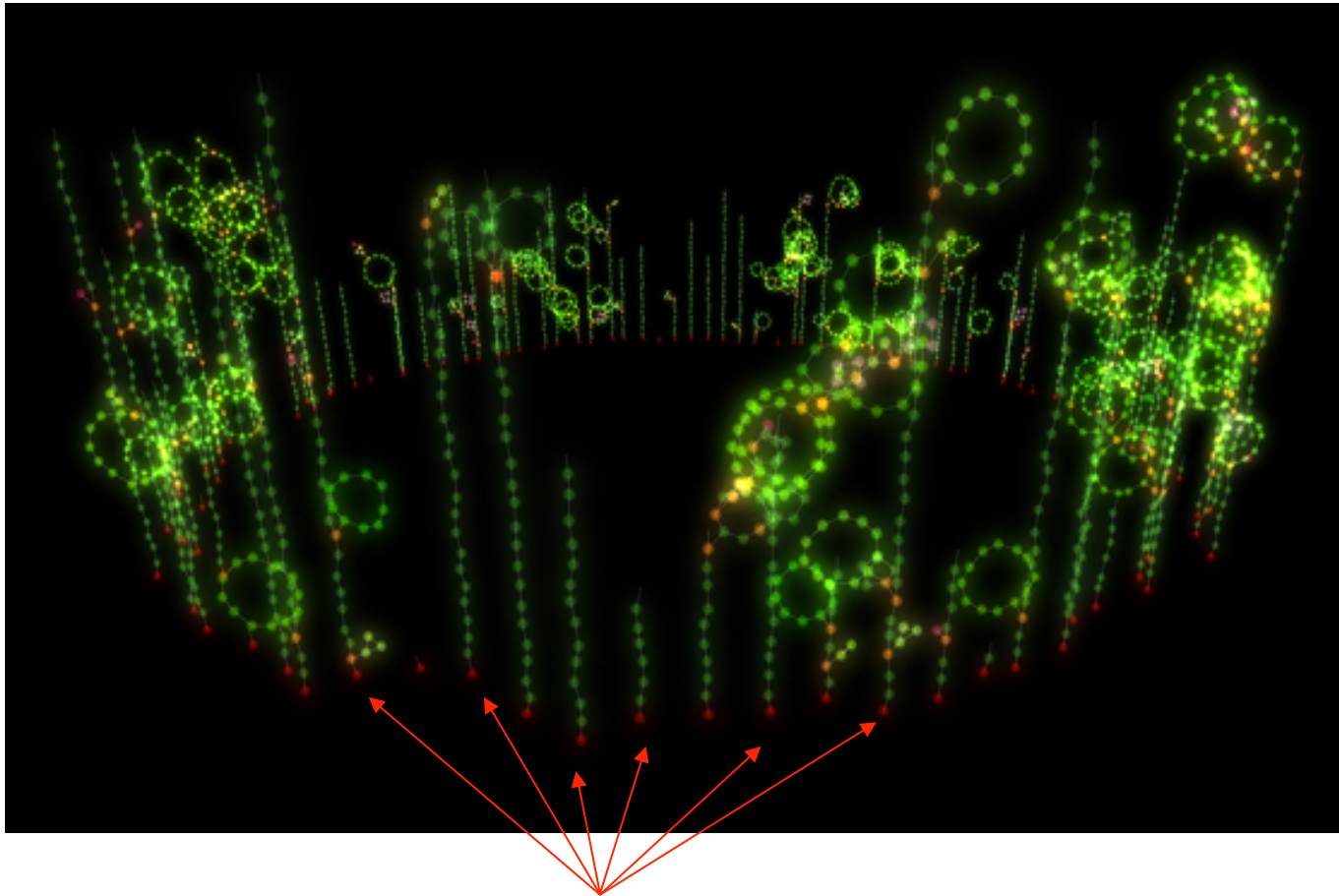
¹ Computer Science Dep., University of Pisa, Italy {rossetti,pedre}@di.unipi.it

² ISTI - CNR KDDLab, Pisa, Italy {fosca.giannotti, giulio.rossetti}@isti.cnr.it

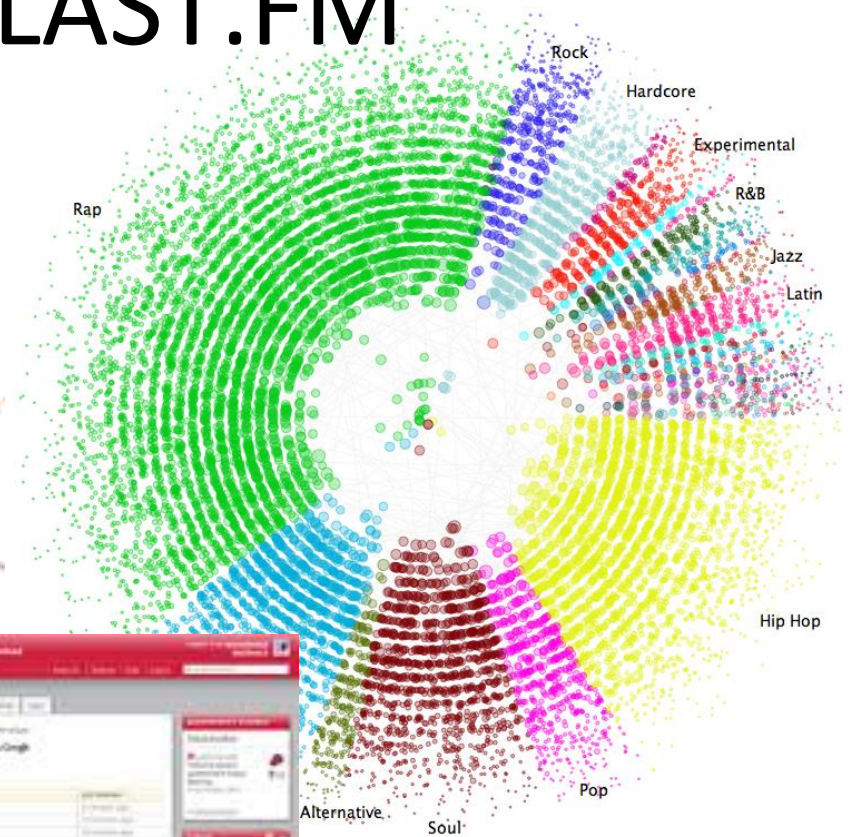
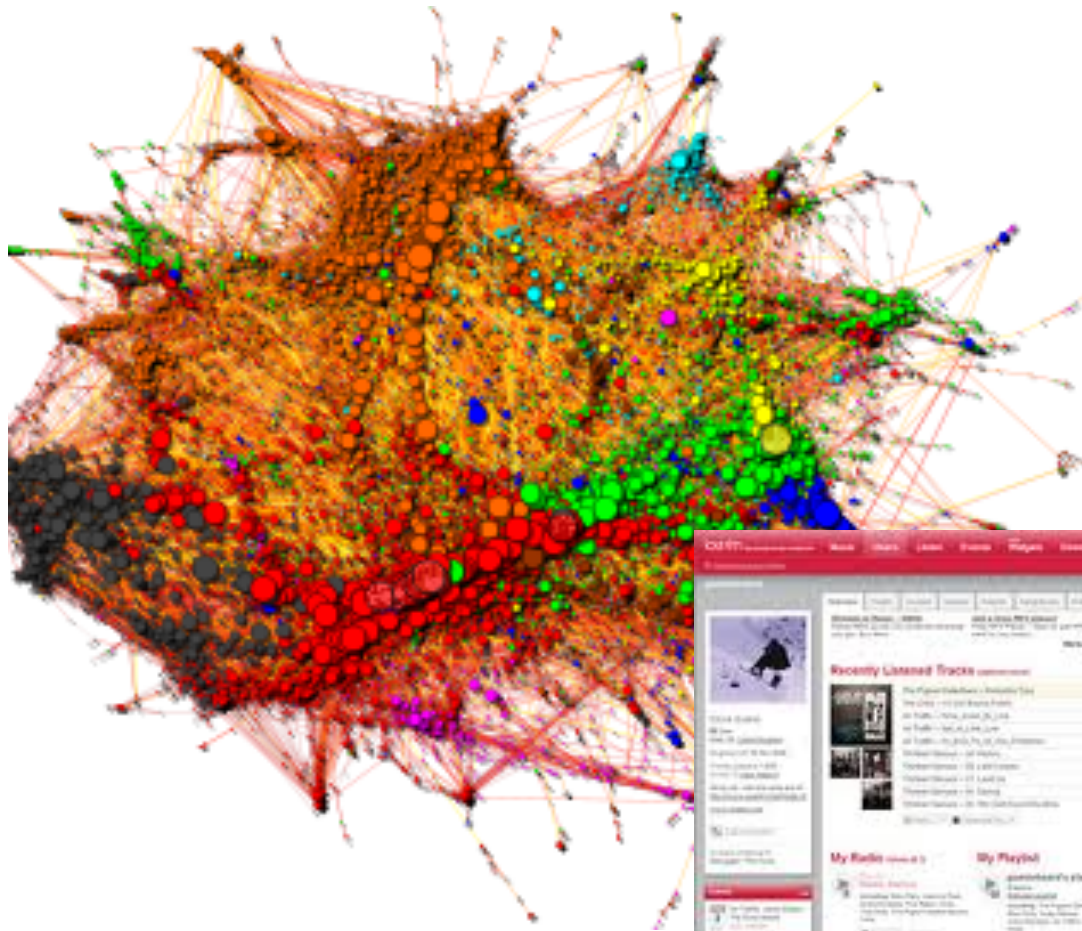
³ IMT Lucca, Italy



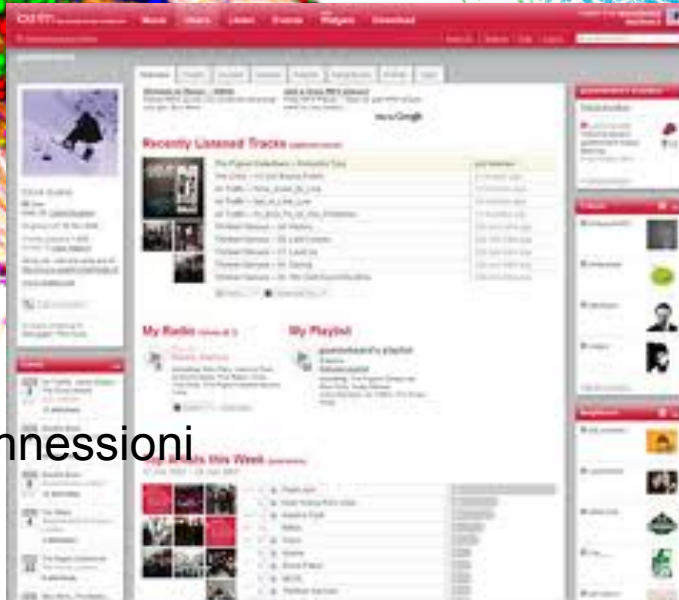
Social Influence: Leaders



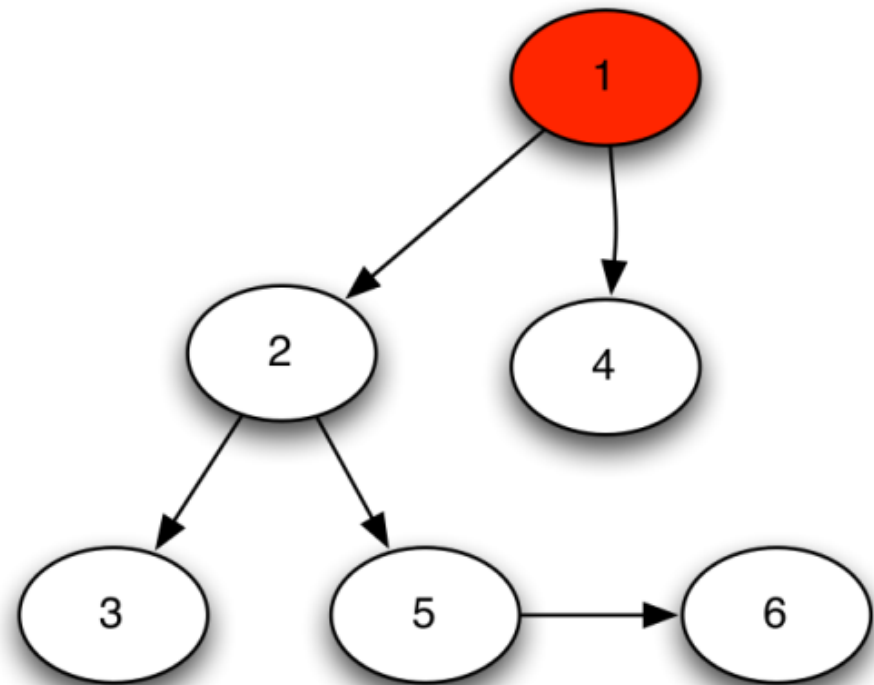
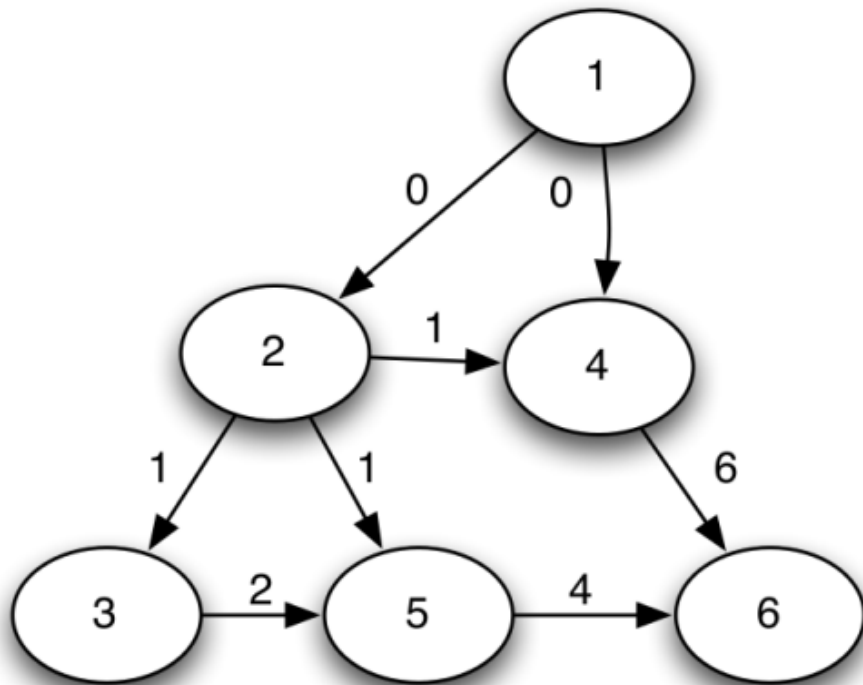
Chiediamo a LAST.FM



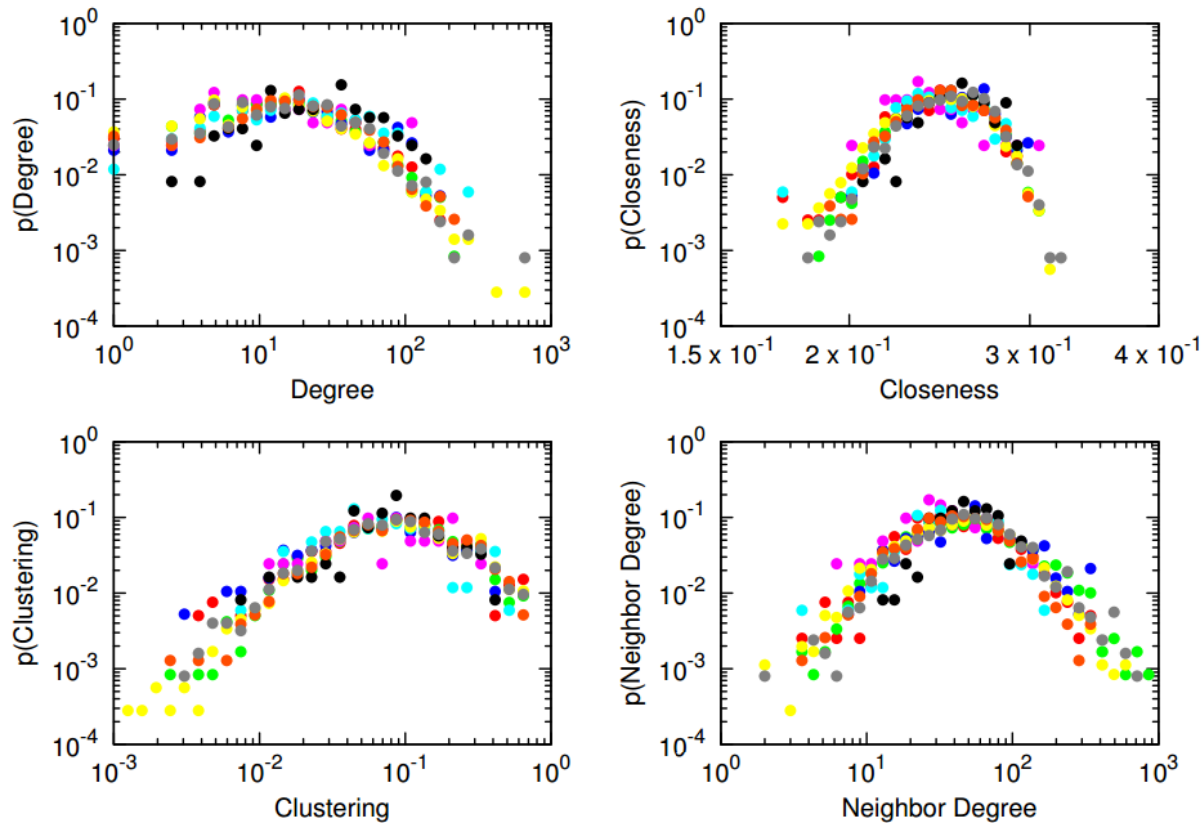
80.000 utenti, 4000.000 connessioni



Leader finding



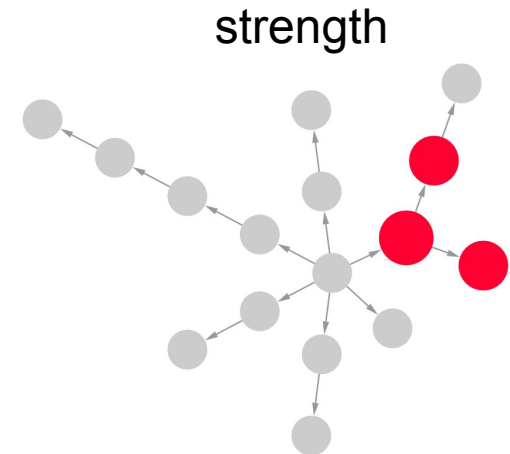
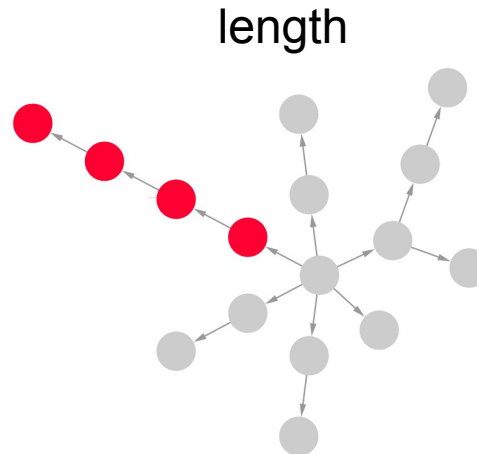
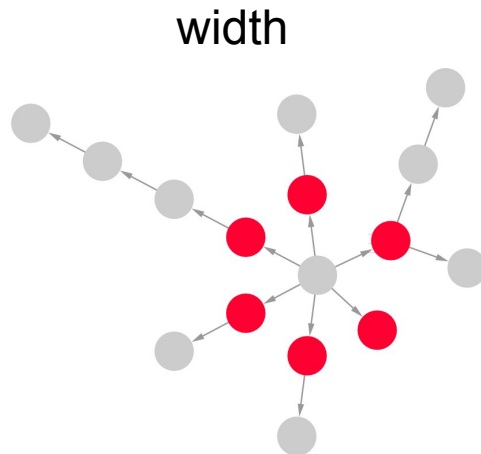
dai BigData...i veri influenzzer non sono i leaders



... abbiamo scoperto che i leader teorici, quelli che avrebbero in teoria il potere di influenzare la rete sociale, non hanno una grande influenza pratica sulla rete.

What is Social Prominence?

- It has been observed that a small set of users in a Social Network is able to anticipate (or influence) the behavior of the entire network
- We detected 3 possible scenarios:

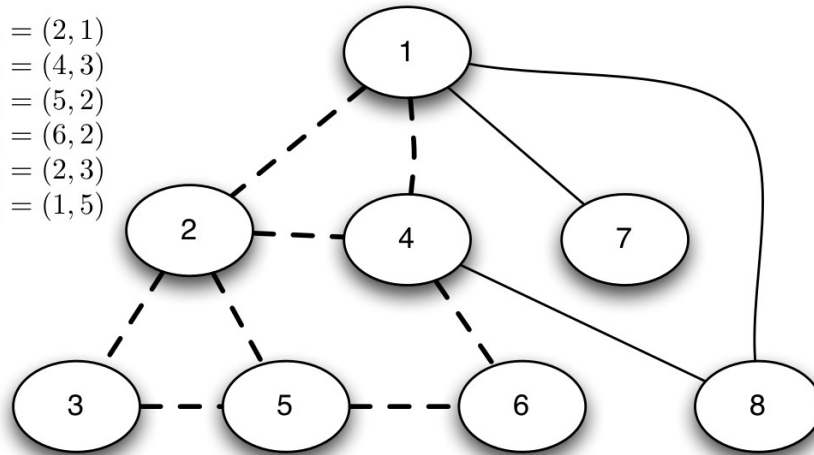


The Idea

- Define what a “leader” is
- Identify three measures of social prominence (width, depth and strength)
- Analyze their relationship with the topological characteristic of prominent actors in a network
- Look for patterns distinguishing different objects spreading in a social network

Leaders and structure

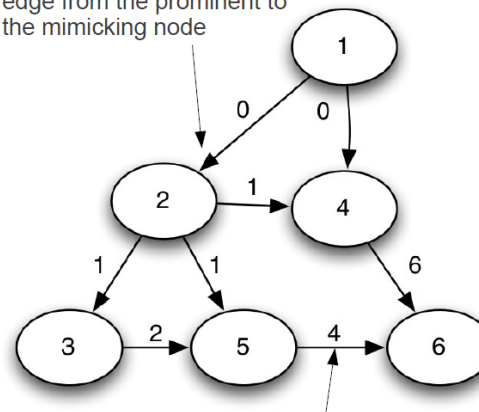
$a_{1,x} = (1, 0)$ $a_{4,y} = (2, 1)$
 $a_{2,x} = (2, 1)$ $a_{7,y} = (4, 3)$
 $a_{3,x} = (1, 2)$ $a_{8,y} = (5, 2)$
 $a_{4,x} = (4, 6)$ $a_{6,y} = (6, 2)$
 $a_{5,x} = (1, 4)$ $a_{1,y} = (2, 3)$
 $a_{6,x} = (6, 7)$ $a_{2,y} = (1, 5)$



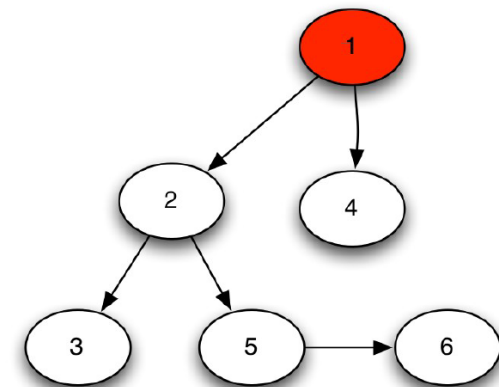
For each Artist we extract the induced temporal subgraph of its Listeners

We define Leader all those nodes that are the first, in their neighborhood to adopt the given artist

Each social connection is transformed in a directed edge from the prominent to the mimicking node



The label on the edge represents the timestep in which the prominent node performed the action



The Minimum Diffusion Tree (MDT) is then the minimum spanning tree

Data, experiments and results

Data gently provided by 

	Width	Strength	Degree	Clustering	Neigh Deg	Bet Centr	Clo Centr
AVG Depth	-0.03	-0.23	-0.08	0.05	-0.08	-0.02	-0.13
Width	-	0.01	-0.31	0.13	0.05	-0.07	-0.59
Strength	-	-	0.02	-0.02	0.03	0.00	0.04
Degree	-	-	-	-0.16	-0.02	0.77	0.56
Clustering	-	-	-	-	-0.05	-0.06	-0.32
Neigh Deg	-	-	-	-	-	-0.00	0.39
Bet Centr	-	-	-	-	-	-	0.22

Central nodes are characterized by low Depth & Width

High Width are usually reached only by nodes in tightly knit communities

There is a trade-off between Depth and Strength (not between D and W nor between S &

Data, experiments and results

Cluster	size	dance	ele	folk	jazz	met	pop	punk	rap	rock
0	1822	1.25	1.13	1.54	1.37	1.50	0.76	1.31	1.13	1.10
1	136	1.28	1.55	1.28	2.35	0.78	0.73	0.64	1.35	0.70
2	664	0.59	0.87	0.98	0.48	0.95	0.97	1.50	1.20	1.19
3	482	1.26	1.16	1.09	1.12	0.91	0.80	2.48	1.24	0.89
4	973	1.14	1.20	1.15	1.41	0.80	0.91	0.66	0.97	0.97
5	512	1.29	0.96	0.95	1.09	1.10	0.97	0.33	1.06	1.01
6	682	0.89	0.79	0.61	0.64	1.13	1.08	1.07	1.08	1.01
7	124	0.75	1.45	0.35	0.64	0	1.09	0	1.02	0.62
8	524	0.93	1.01	1.12	0.91	1.15	1.07	0.43	0.95	0.87
9	937	0.40	0.46	0.19	0.23	0.45	1.56	0.13	0.37	1.06
10	232	0.72	0.57	0.27	0.99	0.38	1.44	0.38	0.46	1.00
11	612	0.74	0.94	0.71	0.40	0.70	1.27	0.07	0.68	0.83

Jazz:

1 lowest width

4 lowest strength

Not easy to be prominent

Pop:

9, 10, 11

Lowest depth, highest strength

Leaders for pop artists are embedded in groups of users very engaged with the new artist, but not prominent among their friends

Punk:

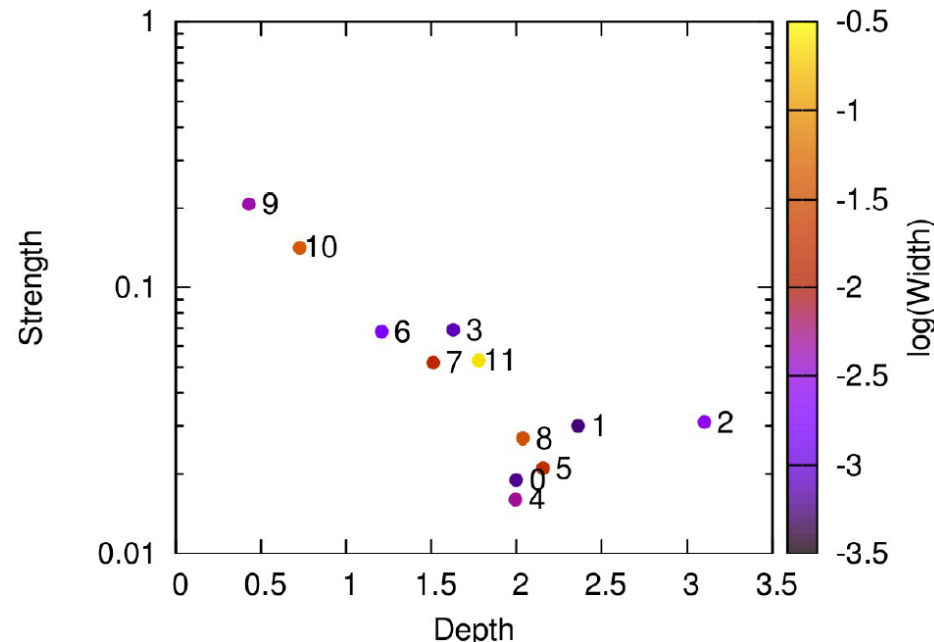
2 high depth, low width and strength

Long cascades, exactly the opposite of the pop genre, similar to folk!

Dance:

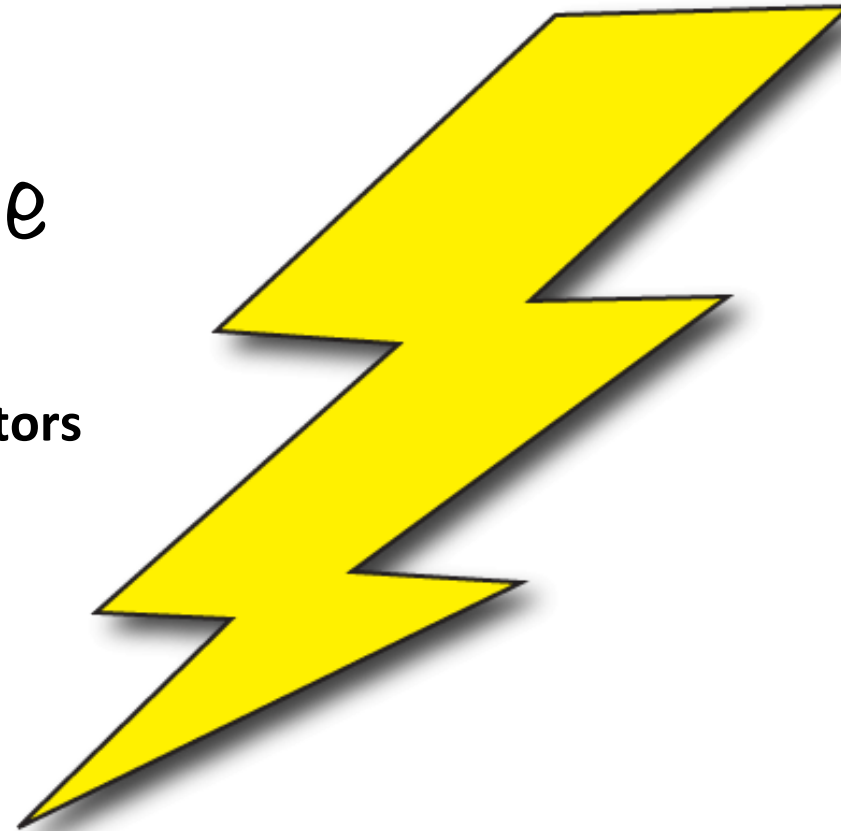
5 high depth, high width, low strength

Dance successes are studied to reach everyone, but in two days nobody remembers anything about them...



“It’s a long way to the
top...”

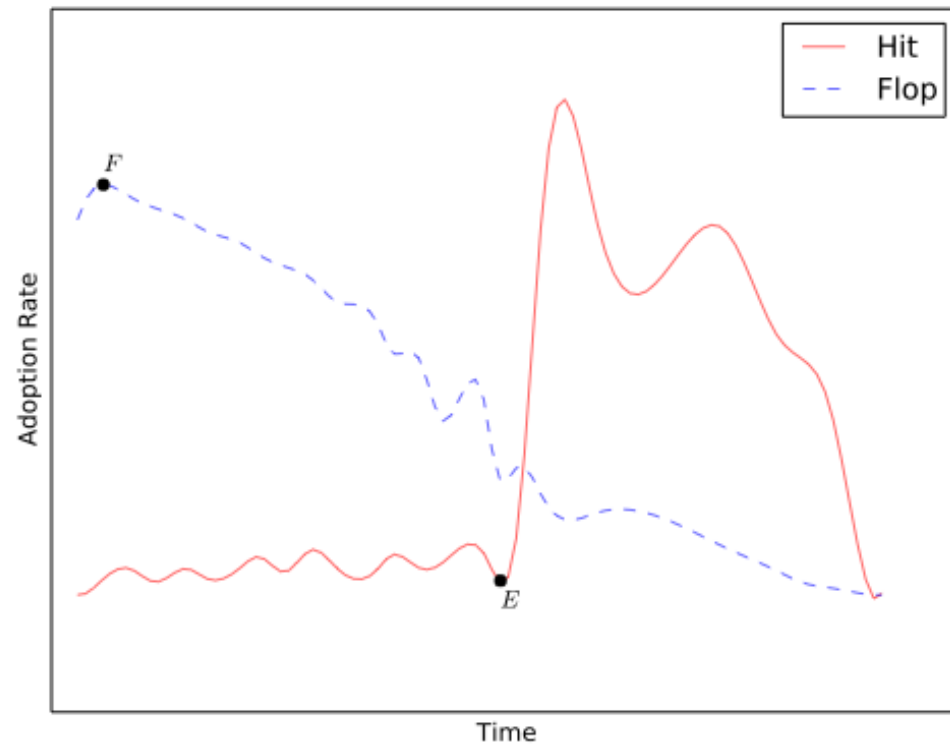
Predicting **Success** via **Innovators**



- G. Rossetti, D. Pennacchioli, L. Milli, D. Pedreschi and F. Giannotti - 2015

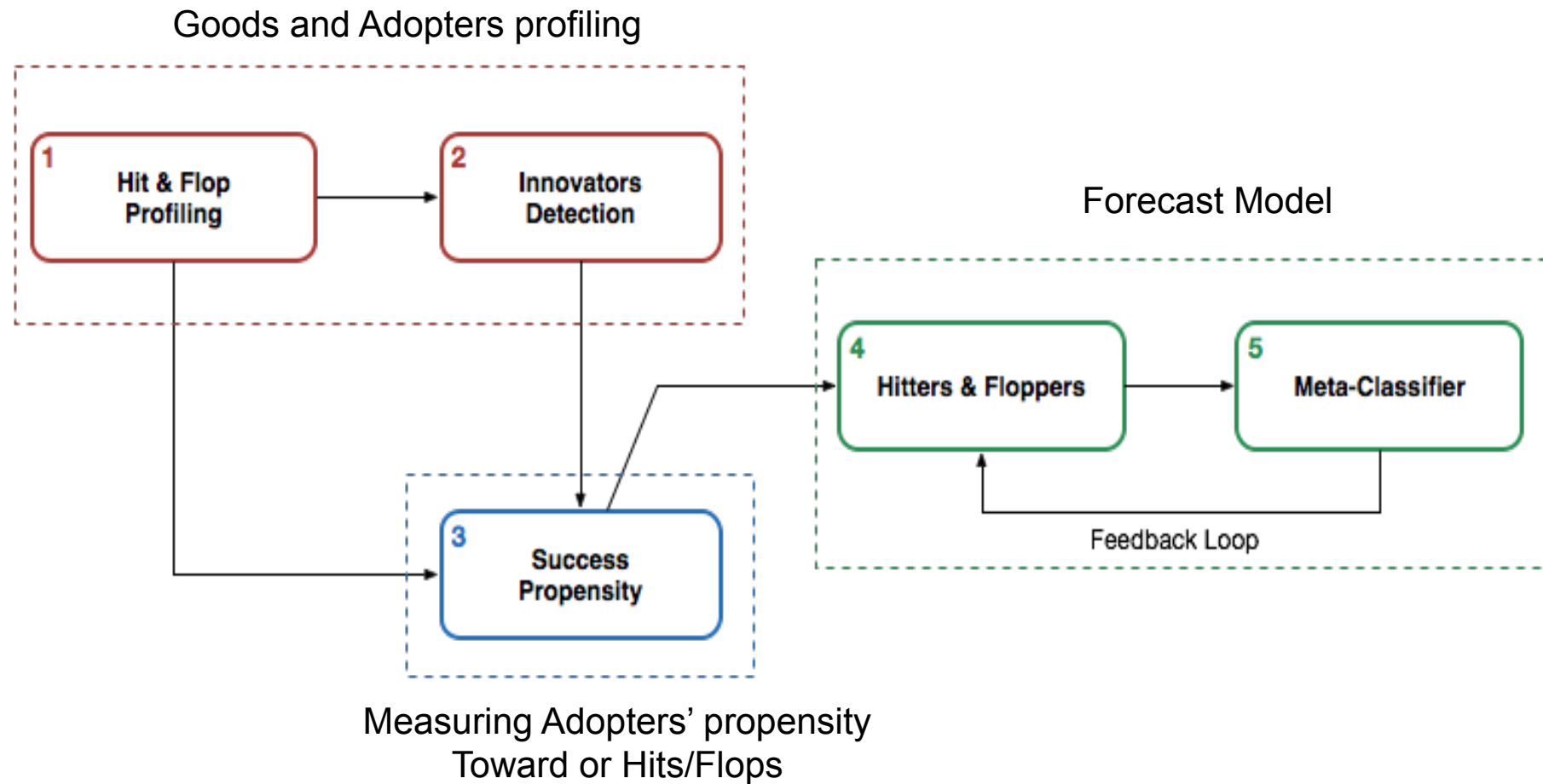
Hits & Flops: qualitative definitions

- **Hit**
 - A good whose trend slowly increases trough time until reaching an explosion point that marks the start of a sharp rising of its adoptions.
- **Flop**
 - A good whose adoption trend does not increase considerably over time or even reaches an early maximum only to sharply decrease.



Given a **partial observation** of the **adoptions** of a **novel good** can we decide if it will become a **Hit** or a **Flop**?

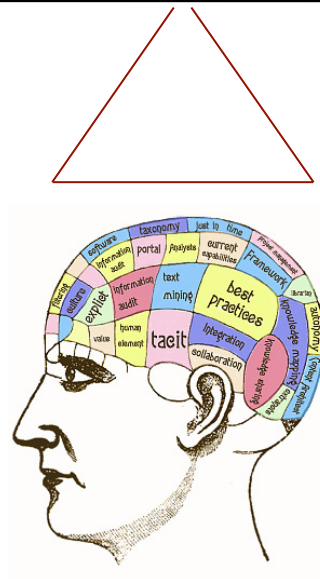
Hit&Flop: Workflow



La Conoscenza Sociale

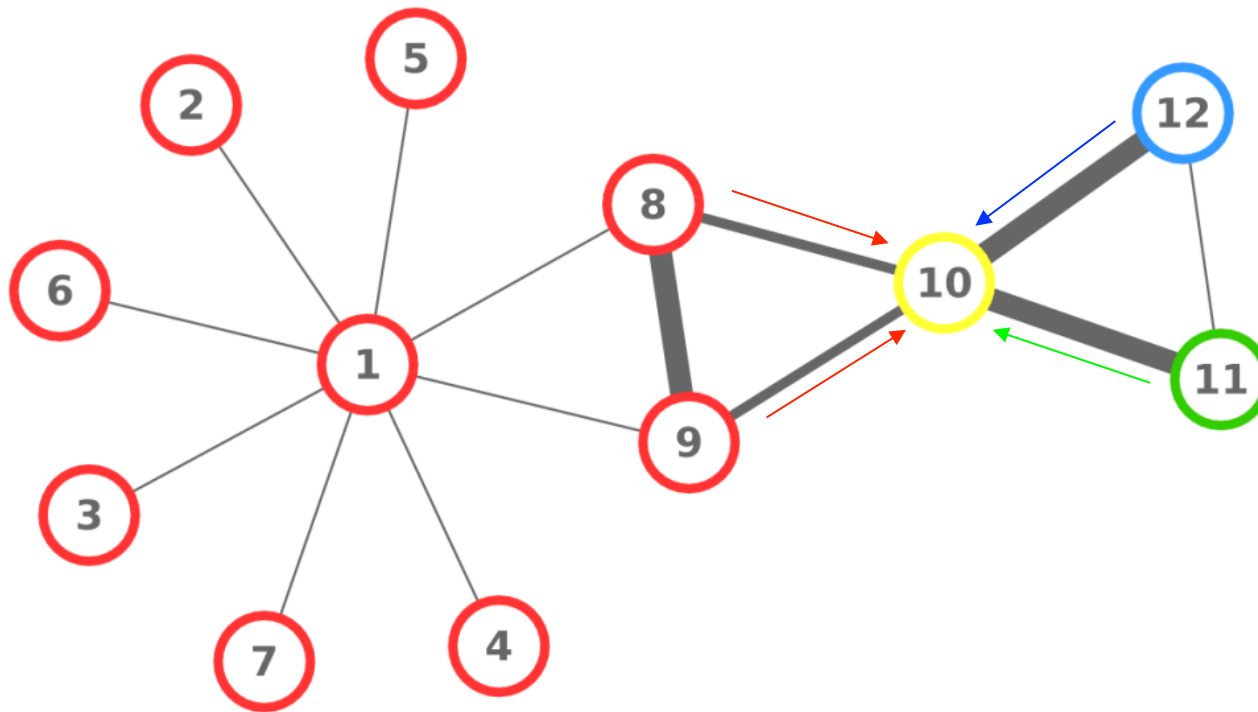


Quello che possiamo
Contenere nel nostro cervello
E' una minima frazione
Della conoscenza umana



Ma le nostre connessioni sociali
Contengono ognuna un'altra parte
E la loro somma puo' essere significat
Come valutare, quindi, una persona?

Calcolare la propria conoscenza sociale



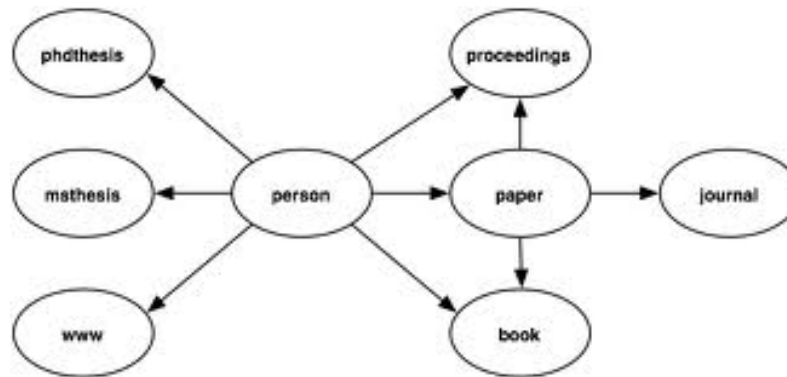
Con processi di diffusione su reti possiamo quantificare l'ammontare di “skill” che ogni connessione ci permette di accedere



dblp

computer science bibliography

Le pubblicazioni di 40.000 ricercatori in DataBase & DataMining per 30 anni



Co-author Graph



Decision-based diffusion models

Herding

[Banerjee 2002]

Herding

- **Influence of actions of others**

- Model where everyone sees everyone else's behavior

- **Sequential decision making**

- **Example: Picking a restaurant**

- Consider you are choosing a restaurant in an unfamiliar town
- Based on Yelp reviews you intend to go to restaurant A
- But then you arrive there is no one eating at A but the next door restaurant B is nearly full

- **What will you do?**

- Information that you can infer from other's choices may be more powerful than your own

■ Herding:

- There is a decision to be made
- People make the decision sequentially
- Each person has some private information that helps guide the decision
- You can't directly observe private information of the others but can see what they do
 - **You can make inferences about the private information of others**

Herding experiment

- Consider an urn with 3 marbles. It can be either:
 - **Majority-blue**: 2 blue, 1 red, or
 - **Majority-red**: 1 blue, 2 red
- Each person wants to **best guess** whether the urn is **majority-blue** or **majority-red**
 - Guess **red** if $P(\text{majority-red} \mid \text{what she has seen or heard}) > \frac{1}{2}$
- **Experiment**: One by one each person:
 - Draws a marble
 - **Privately** looks at the color and puts the marble back
 - **Publicly** guesses whether the urn is **majority-red** or **majority-blue**
- You see all the guesses beforehand.
How should you make your guess?

- **State of the world:**
 - Whether the urn is **MR** or **MB**
- **Payoffs:**
 - Utility of making a correct guess
- **Signals:**
 - Models private information:
 - The color of the marble that you just draw
 - Models public information:
 - The **MR** vs **MB** guesses of people before you

Sequential decision making

- **Decision:** Guess **MR** if $P(\text{MR} | \text{past actions}) > \frac{1}{2}$
- **Analysis (Bayes rule):**
 - #1 follows her own color (private signal)!
 - Why?
$$P(MR | r) = \frac{P(MR)P(r | MR)}{P(r)} = \frac{1/2 \cdot 2/3}{1/2} = 2/3$$
$$P(r) = P(r | MB)P(MB) + P(r | MR)P(MR) = \frac{1}{2} \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{2}{3} = 1/2$$
 - #2 guesses her own color (private signal)!
 - #2 knows #1 revealed her color. So, #2 gets 2 colors.
 - If they are the same, decision is easy.
 - If not, break the tie in favor of her own color

Sequential decision making

- #3 follows majority signal!

- Knows #1, #2 acted on their colors. So, #3 gets 3 signals.
- If #1 and #2 made opposite decisions, #3 goes with her own color. Future people will know #3 revealed its signal

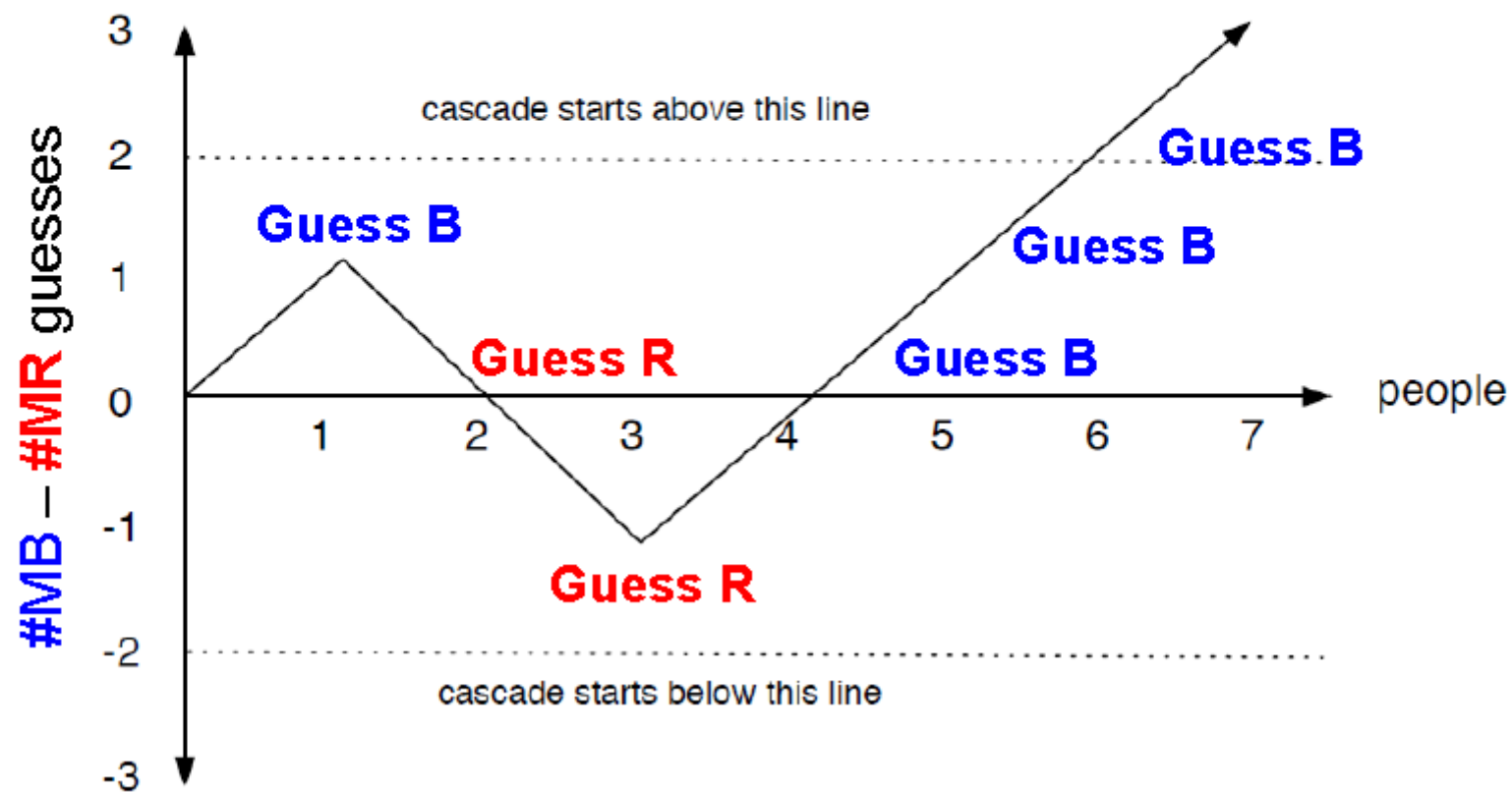
$$P(MR | r, r, b) = 2/3$$

- If #1 and #2 made same choice, #3's decision conveyed no info. **Cascade has started!**

- How does this unfold? You are N-th person

- #MB = #MR : you guess your color
- |#MB - #MR| = 1 : your color makes you indifferent, or reinforces you guess
- |#MB - #MR| ≥ 2 : Ignore your signal. Go with majority.

- Cascade begins when the difference between the number of blue and red guesses reaches 2



Influence maximization

How to create big cascade?

- **Blogs – Information epidemics**

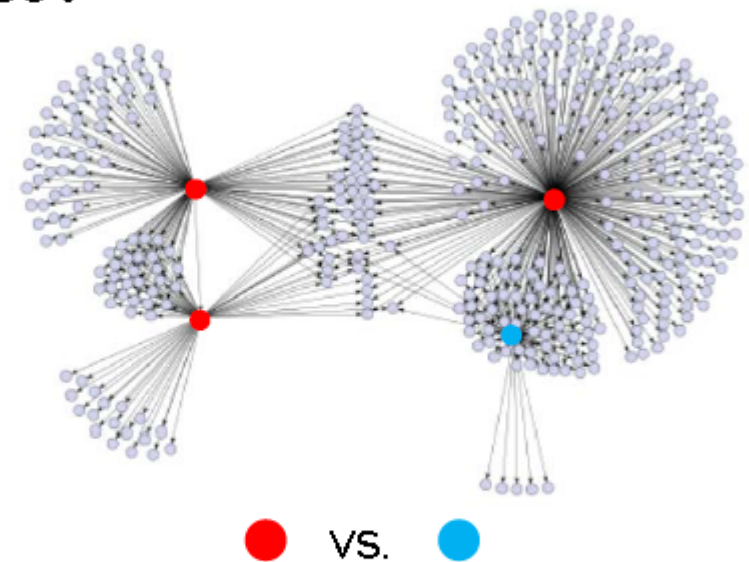
- Which are the influential/infectious blogs?
- Which blogs create big cascades?

- **Viral marketing**

- Who are the influencers?
- Where should I advertise?

- **Disease spreading**

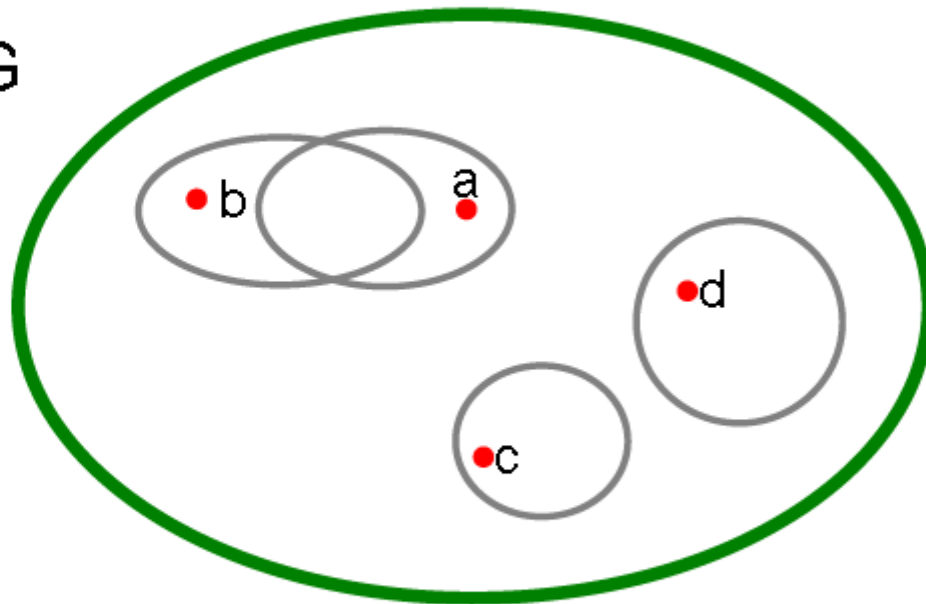
- Where to place monitoring stations to detect epidemics?



Most influential sets of nodes

- **S**: is initial active set
- $f(S)$: the expected size of final active set

graph G



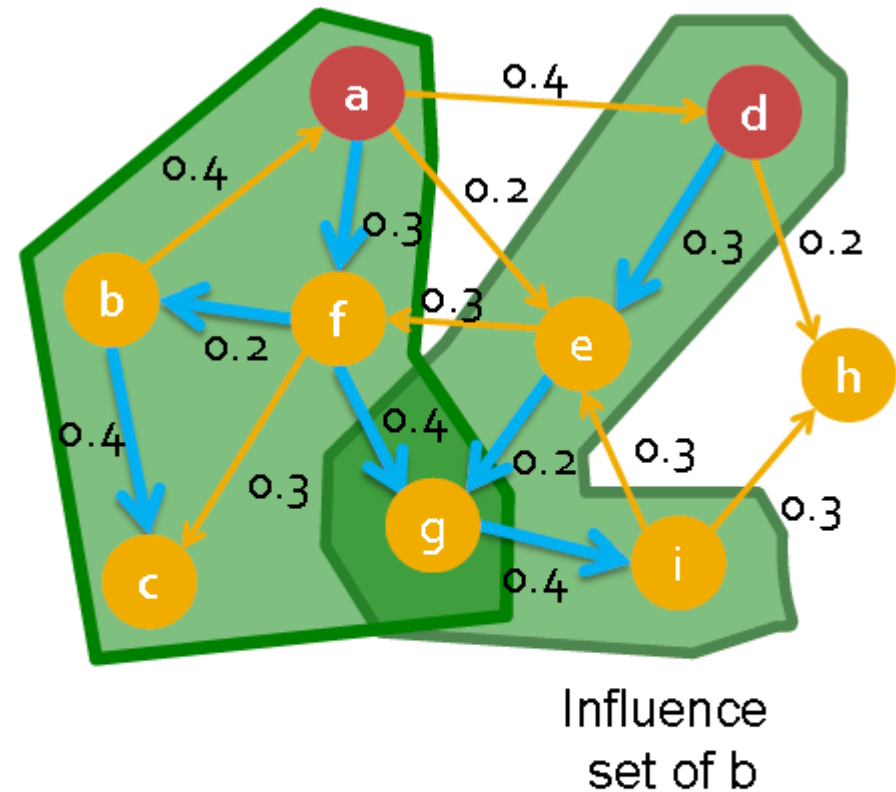
○ ... influence set
of a node

- **Set S is more influential if $f(S)$ is larger**
 $f(\{a,b\}) < f(\{a,c\}) < f(\{a,d\})$

Most influential sets of nodes

Problem:

- **Most influential set of size k :** set S of k nodes producing largest expected cascade size $f(S)$ if activated
[Domingos-Richardson '01]



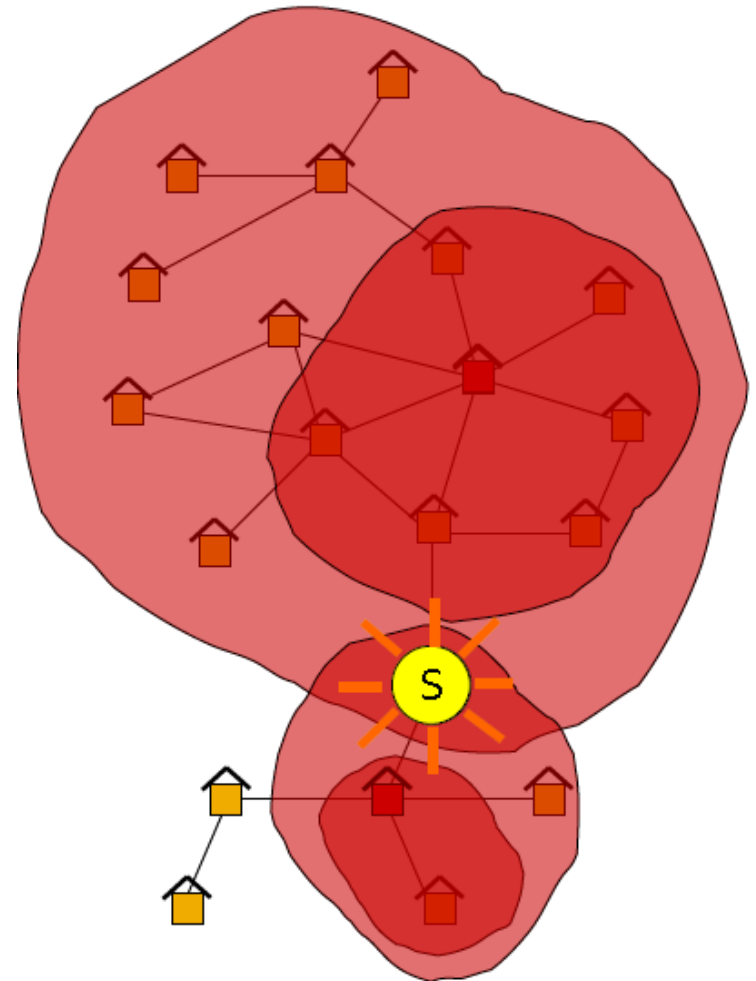
- **Optimization problem:** $\max_{S \text{ of size } k} f(S)$

How hard is the problem?

- NP-HARD!
- But if $f(S)$ is «diminishing returns»
 - Monotonic and submodular
- Then the approximated solution computed with a greedy algorithm (hill climbing) has a bounded distance with the global optimum!

Related problem: outbreak detection

- **Which node(s) initiated a cascade?**
- Given a real city water distribution network
- And data on how contaminants spread in the network
- Detect the contaminant as quickly as possible

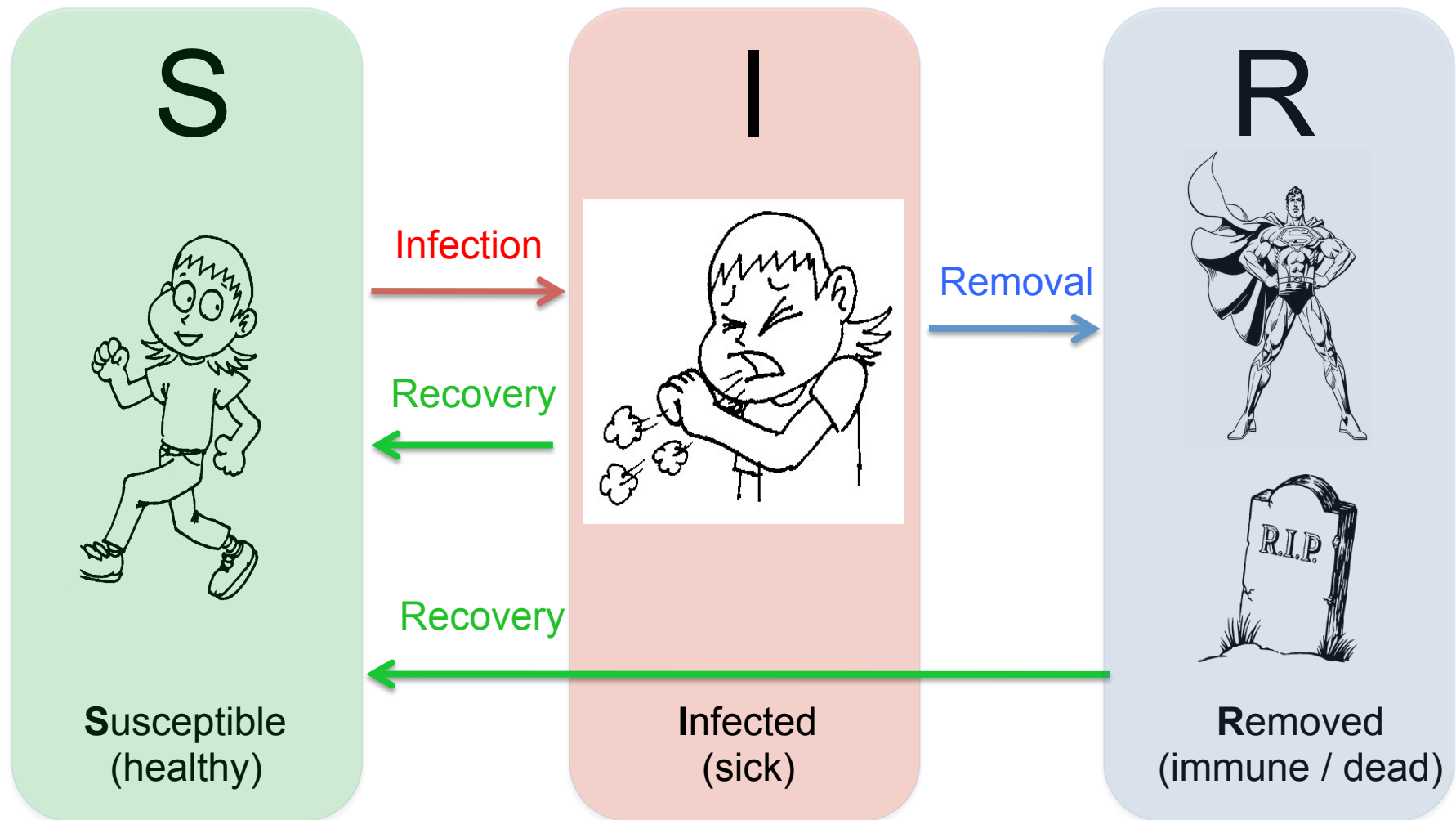


Probabilistic models of diffusion

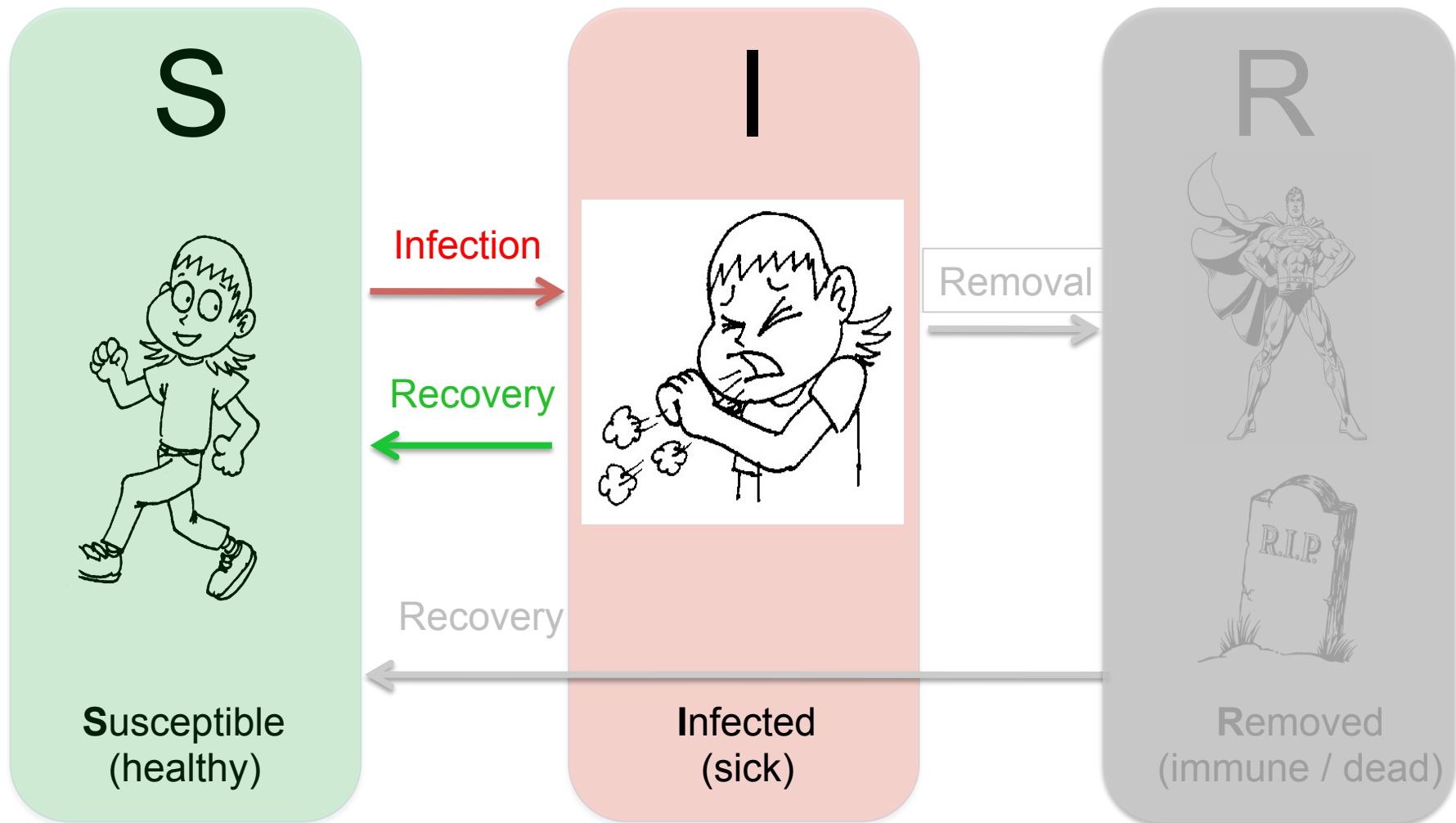
Epidemic modeling

Epidemic Modeling (classical models)

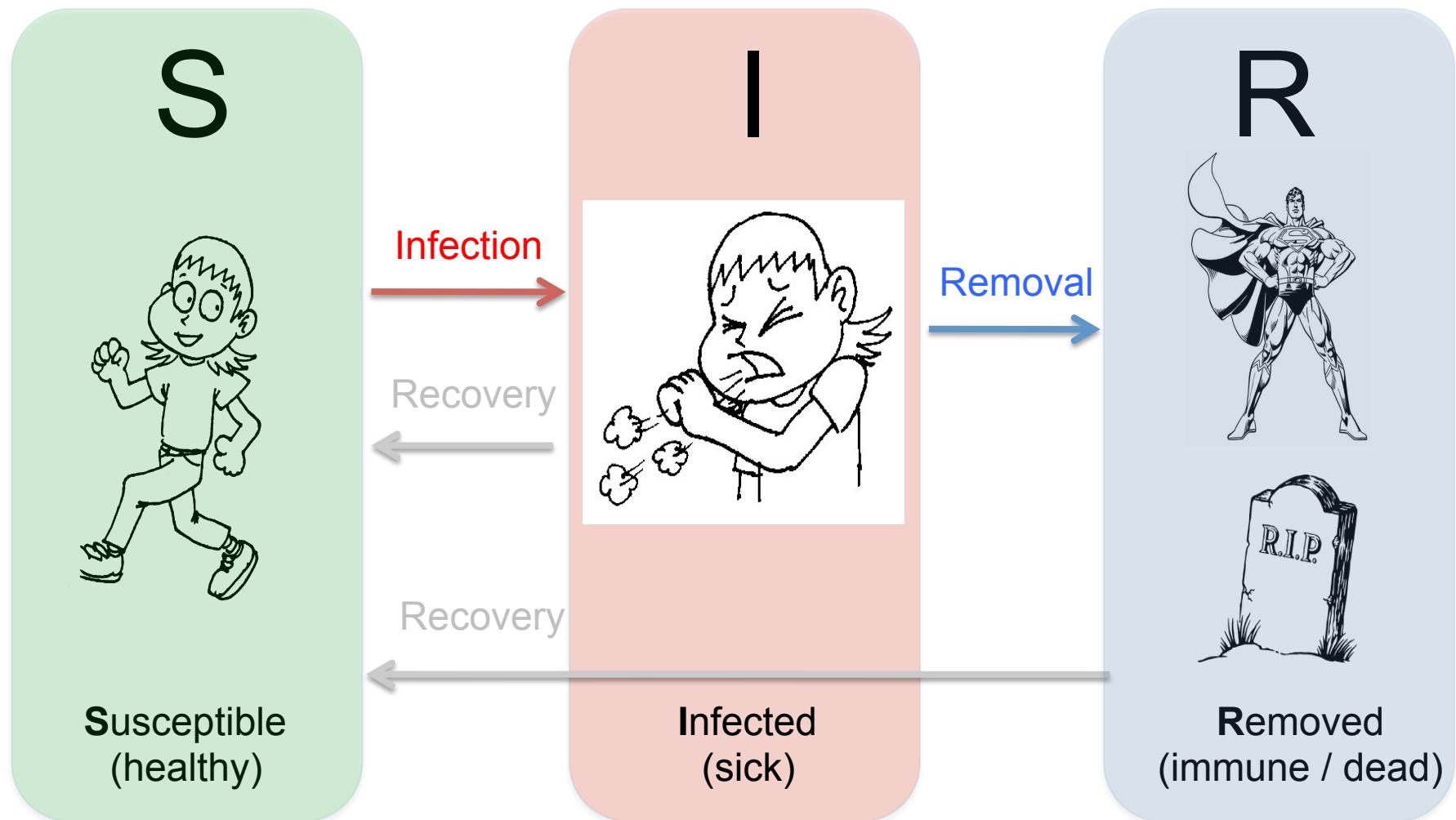
Classical Epidemic Models – Basic States



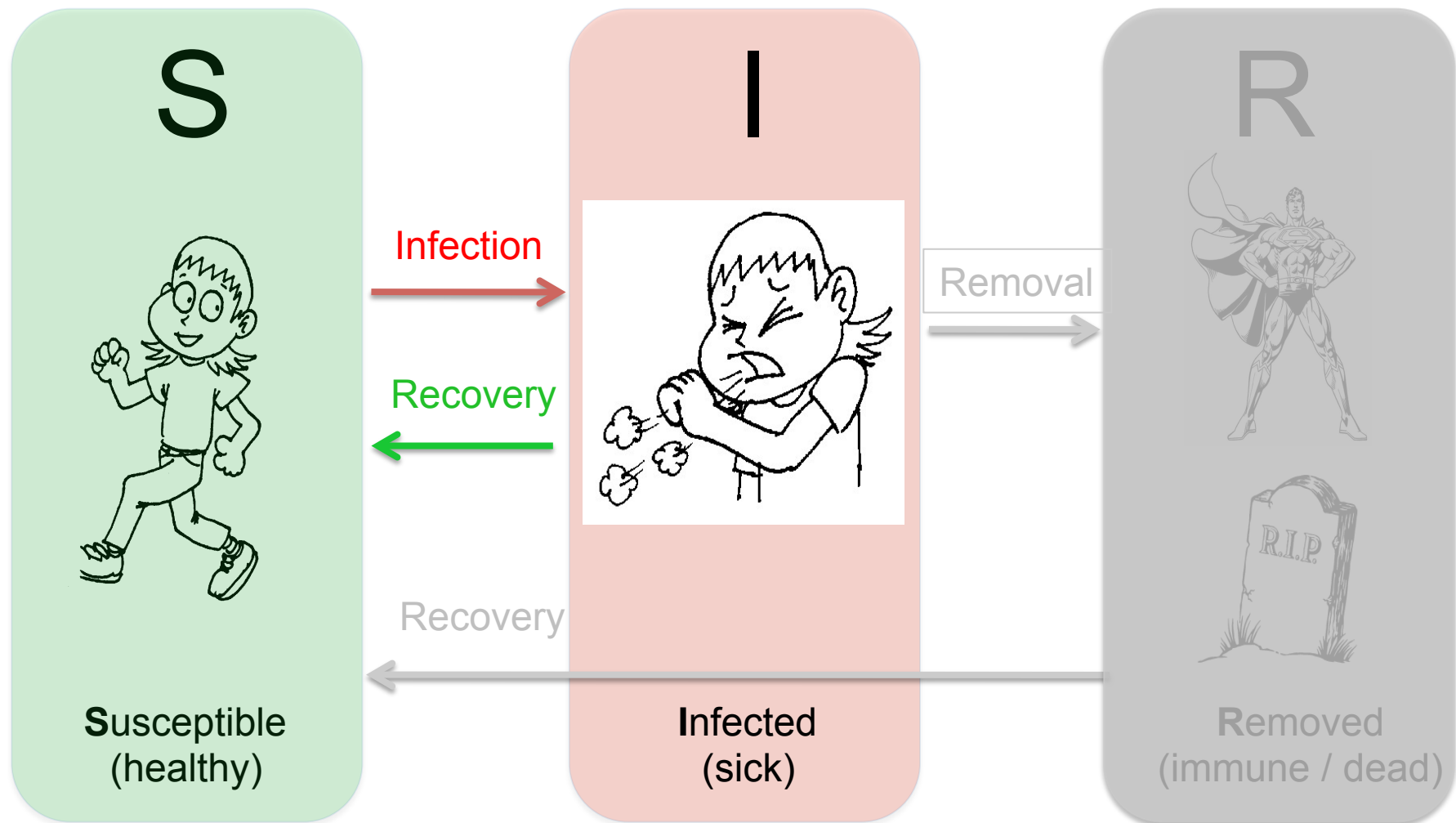
SIS Model: Common Cold



Example 2: Flu, SARS, Plague, ...



SIS Model: Common Cold



SIS Model Dynamics

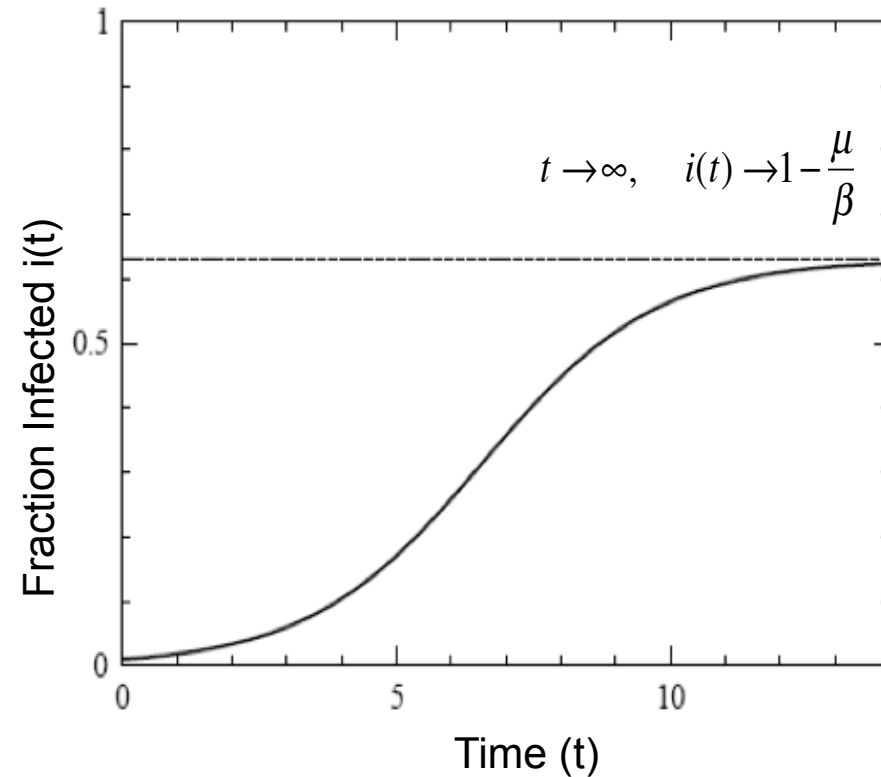
$$\frac{di}{dt} = \underbrace{\beta i}_{I \rightarrow S} (1-i) - \underbrace{\mu i}_{S \rightarrow I} = i(\beta - \mu - \beta i)$$

$$\frac{di}{i} + \frac{di}{1 - \mu/\beta - i} = (\beta - \mu)dt$$

$$\ln(i) - \ln(1 - \mu/\beta - i) = (\beta - \mu)t + c$$

$$\frac{i}{1 - \mu/\beta - i} = Ce^{(\beta - \mu)t} \quad C = e^c$$

$$\therefore i(t) = \left(1 - \frac{\mu}{\beta}\right) \frac{Ce^{(\beta - \mu)t}}{1 + Ce^{(\beta - \mu)t}}$$



Stationary state:

$$\frac{di}{dt} = \beta i(1-i) - \mu i = 0$$

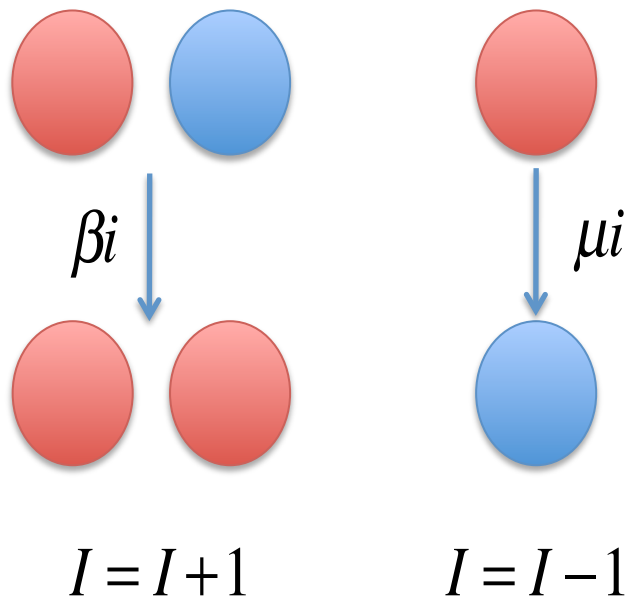
SIS model: fraction infected individuals saturates below 1.

SIS Model: Epidemic Threshold and Basic Reproductive Number

$$\frac{di}{dt} = \underbrace{\beta}_{\text{I}} \underbrace{i(1-i)}_{\text{S}} - \underbrace{\mu i}_{\text{I} \rightarrow \text{S}}$$

If $\mu \approx \beta$, $i \rightarrow 0$

“Epidemic threshold”



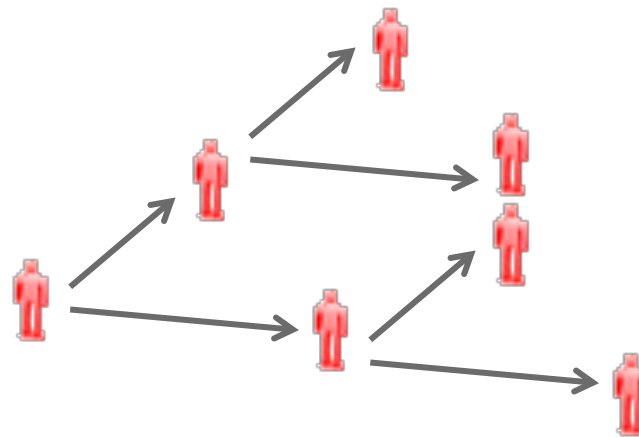
$$\lambda \equiv \frac{\beta}{\mu} \quad \text{“Basic reproductive number”}$$

On average, how many infected individuals will be infected by one infected individual?

$\lambda > 1$: Outbreak, $\lambda < 1$: Die out

reproductive number λ : average # of infectious individuals generated by one infected in a fully susceptible population.

e.g. $\lambda = 2$



Choose
transmission
scenario

☐ mild

$\lambda = 1.5$

☒ medium

$\lambda = 1.9$

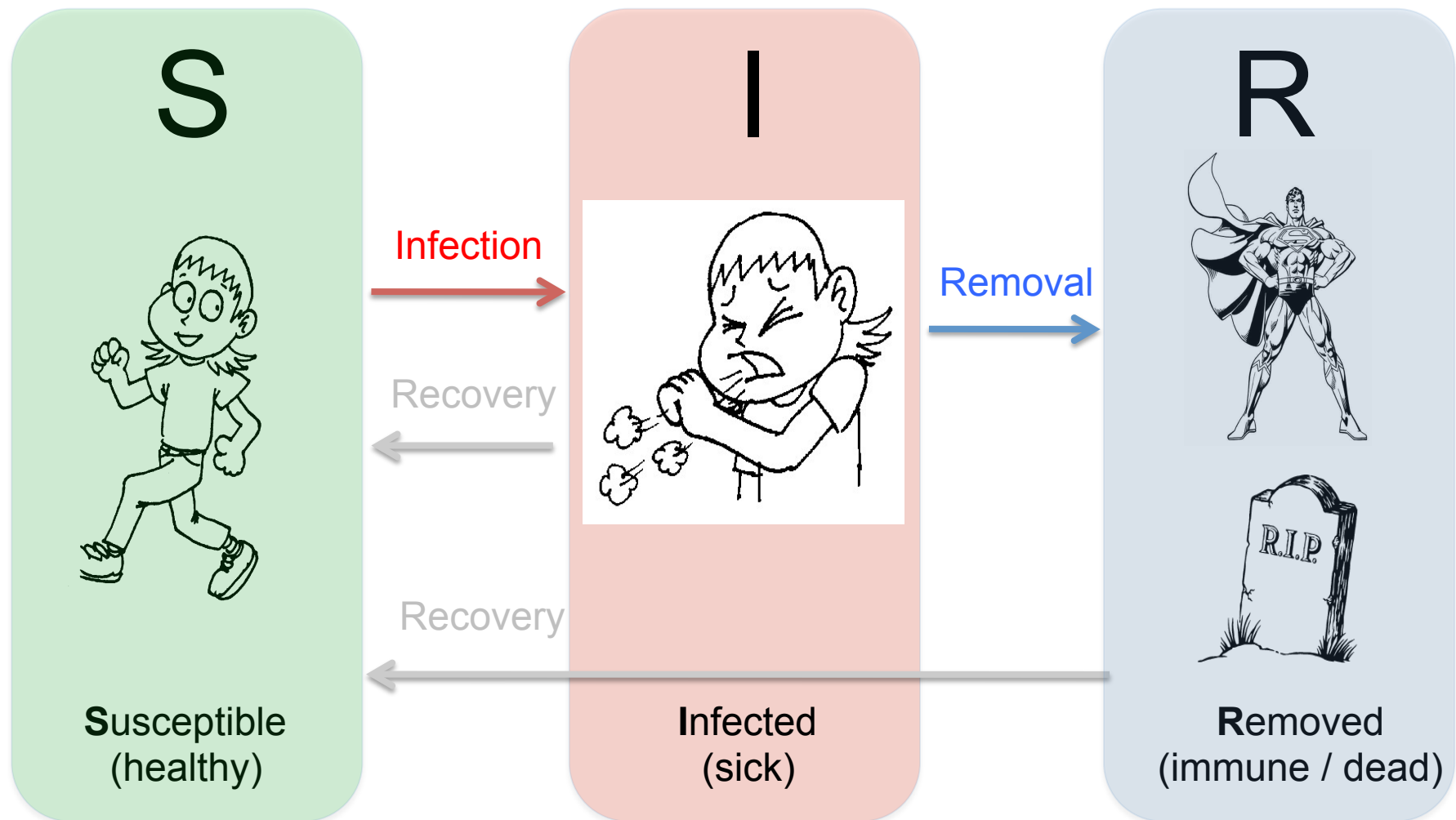
☐ high

$\lambda = 2.3$

☐ very high

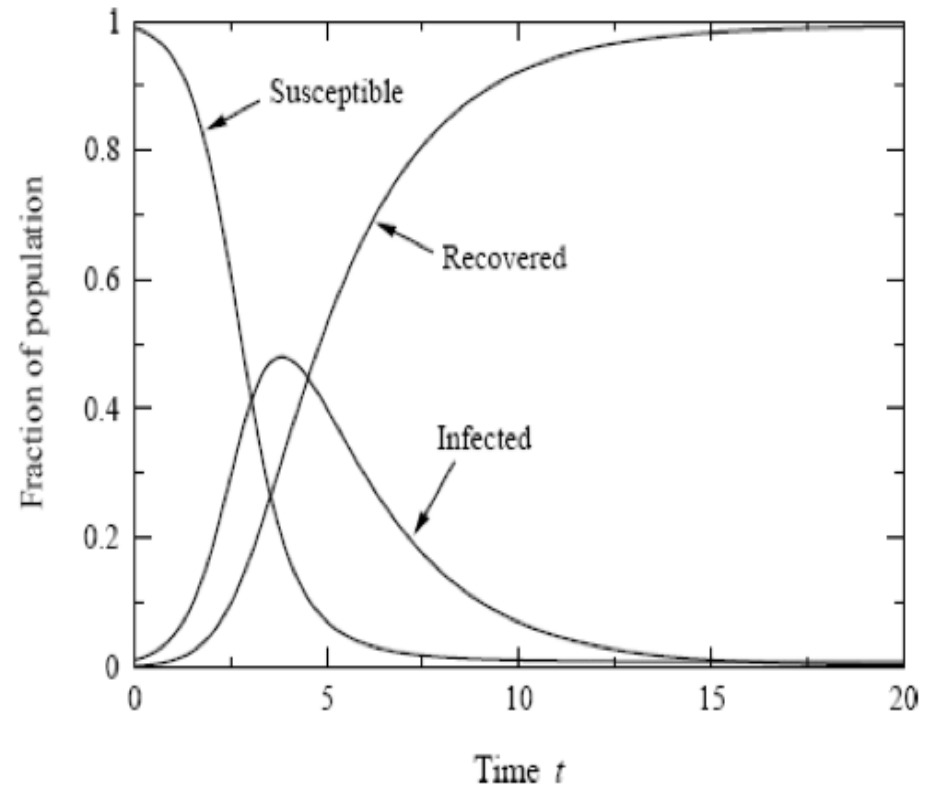
$\lambda = 2.7$

Example 2: Flu, SARS, Plague, ...



SIR Model

$$\begin{aligned}\frac{ds(t)}{dt} &= \beta \langle k \rangle i(t) [1 - r(t) - i(t)] \\ \frac{di(t)}{dt} &= -\mu i(t) + \beta \langle k \rangle i(t) [1 - r(t) - i(t)] \\ \frac{dr(t)}{dt} &= \mu i(t).\end{aligned}$$



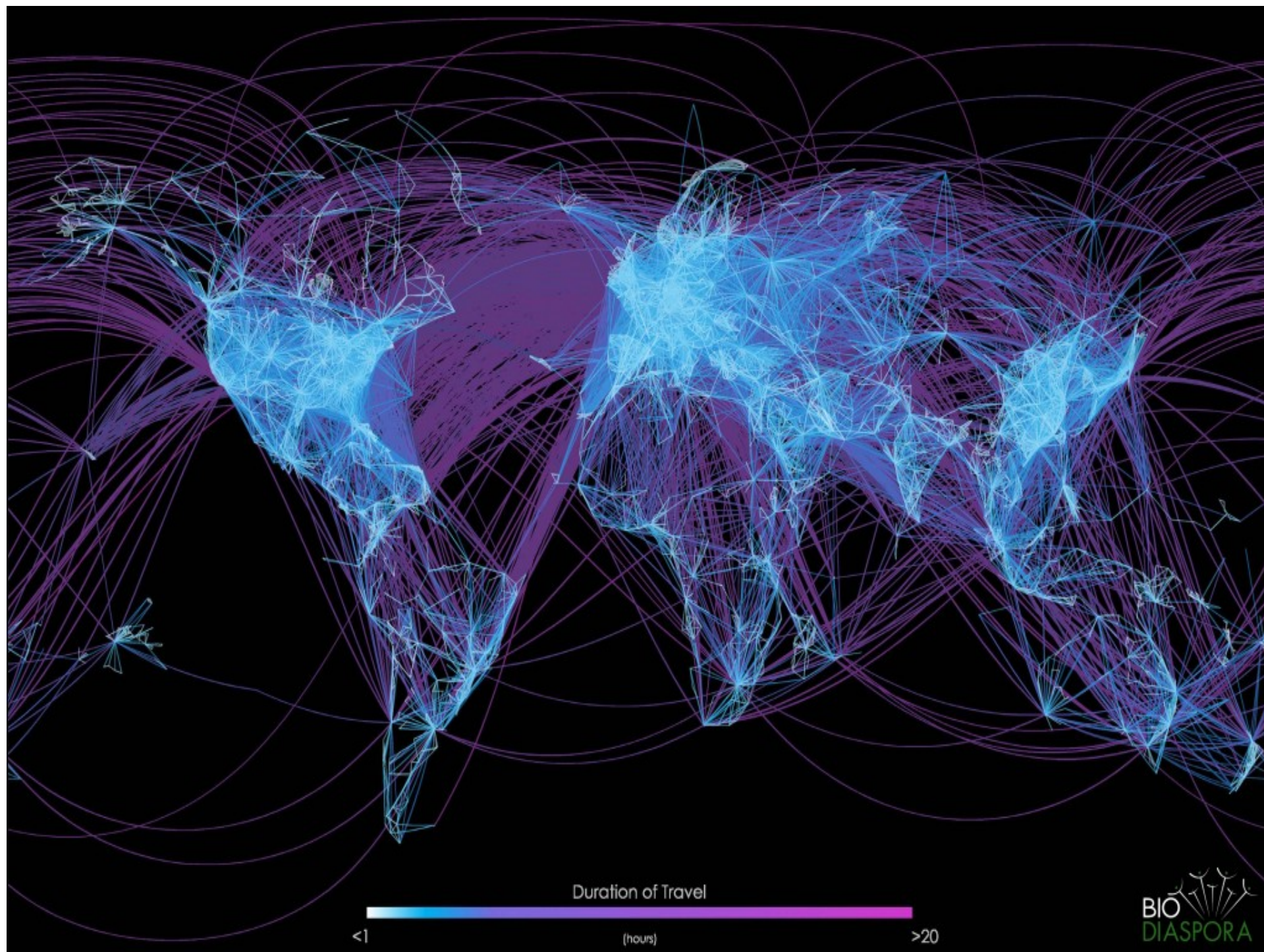
- SIR model: the fraction infected peaks and the fraction recovered saturates.

Epidemic modeling on networks

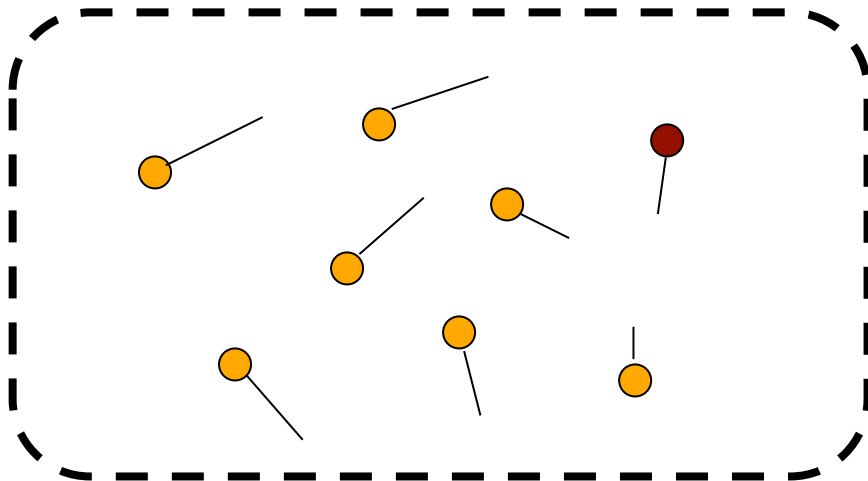
[Vespignani et al., since 2002]

Gleamviz

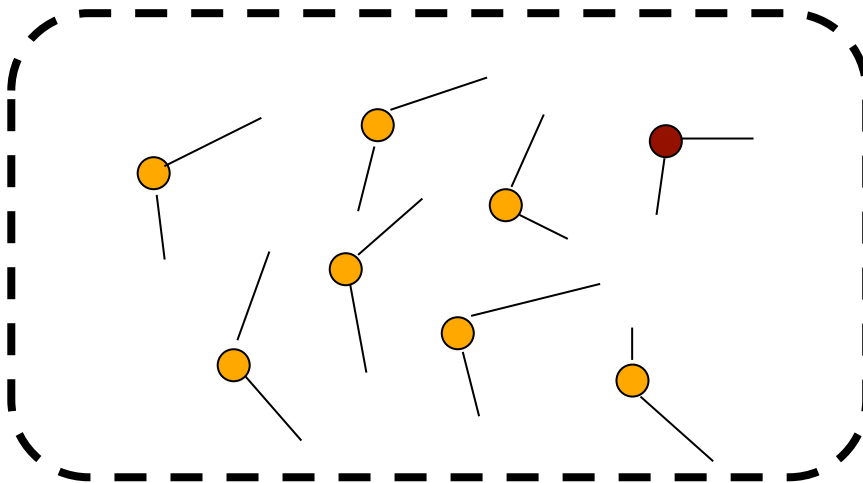




SIS model on a network: Degree based representation



Class of nodes with degree $k=1$



Class of nodes with degree $k=2$

Split nodes by their degrees

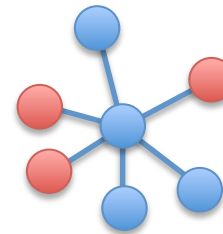
$$i_k = \frac{I_k}{N_k}, \quad i = \sum_k P(k) i_k$$

SIS model:

$$\frac{di_k(t)}{dt} = \beta(1 - i_k(t))k\Theta_k(t) - \mu i_k(t)$$

Proportional to
 k

Density of infected
neighbors of nodes with
degree k



I am susceptible with k
neighbors, and $\Theta_k(t)$
of my neighbors are infected.

(Vespignani)

Early time behavior – SI model – the characteristic time vanishes!

$$\tau = \frac{\langle k \rangle}{\beta(\langle k^2 \rangle - \langle k \rangle)}$$

The timescale it takes for an epidemics to grow. The smaller is τ , the faster it grows.

ER network:

$$\langle k^2 \rangle = \langle k \rangle (\langle k \rangle + 1) \quad \tau_{ER} = \frac{1}{\beta \langle k \rangle}$$

→ The more connected the network is, the faster does the epidemic spread.

SF network ($\gamma < 3$):

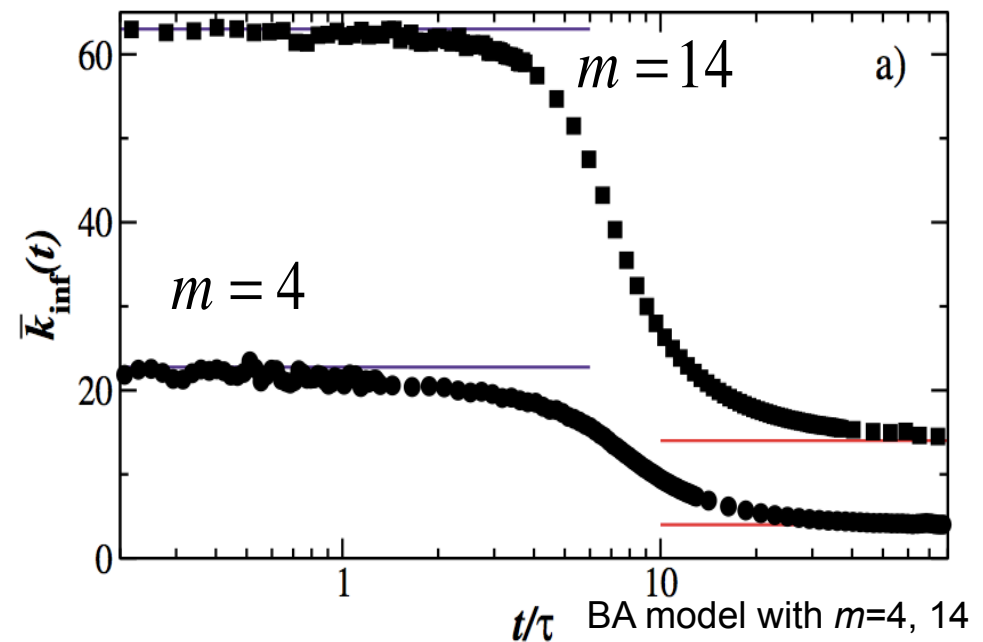
$$\langle k^2 \rangle \rightarrow \infty \text{ for } N \rightarrow \infty \rightarrow \tau \rightarrow 0$$

For heterogeneous networks, the characteristic time vanishes, which means that the epidemic becomes instantaneous. The reason: the hubs get infected first, which then rapidly reach most nodes.

Numerical Test:

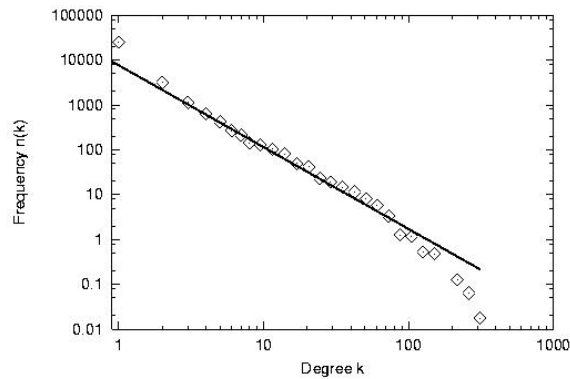
The average degree of newly infected nodes at time t :

$$\bar{k}_{inf}(t) = \frac{\sum_k k(I_k(t) - I_k(t-1))}{I(t) - I(t-1)}$$



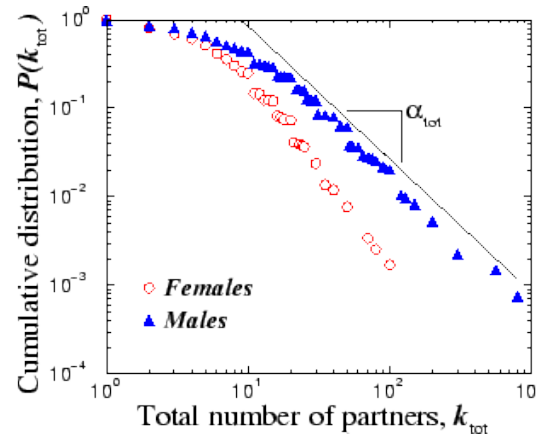
SIS Model – Absence of Epidemic Threshold

Email network



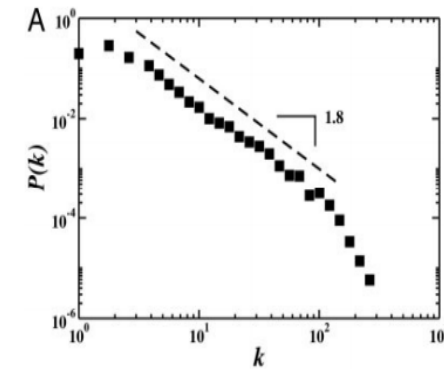
Ebel et al. (2002)

Human sexual network



Liljeros et al., Nature (2001),
Schneeberger et al. STD (2004)

Air transportation network



Colizza et al., PNAS 2006

Many networks will have small or vanishing epidemic threshold!



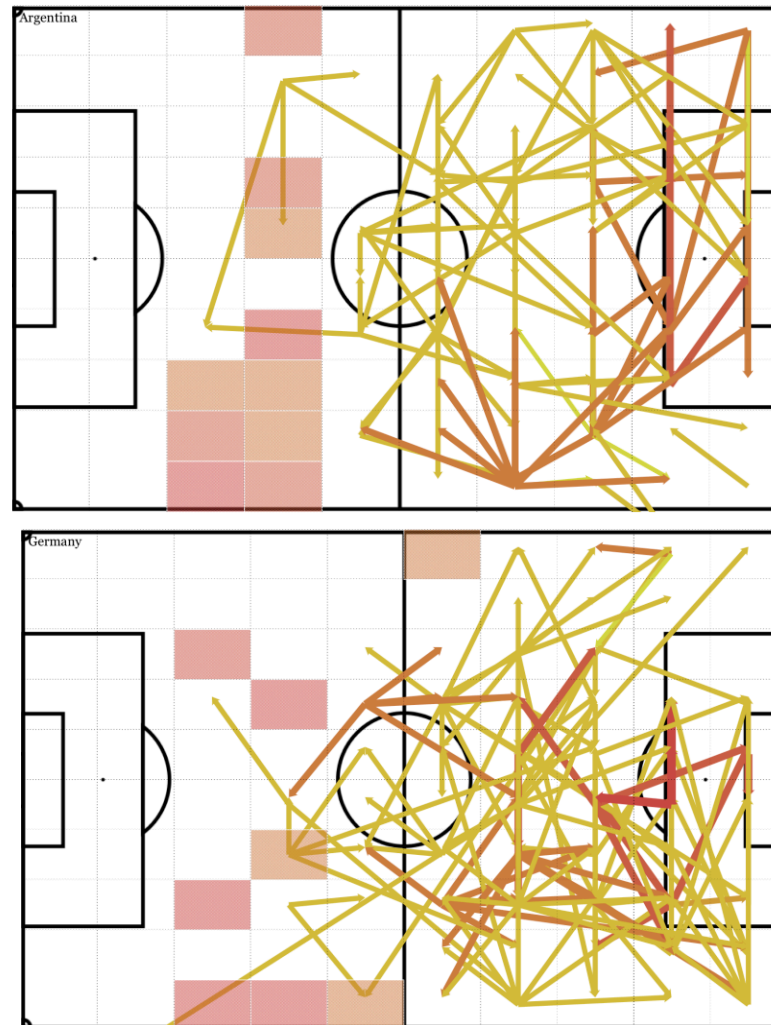
Research highlights

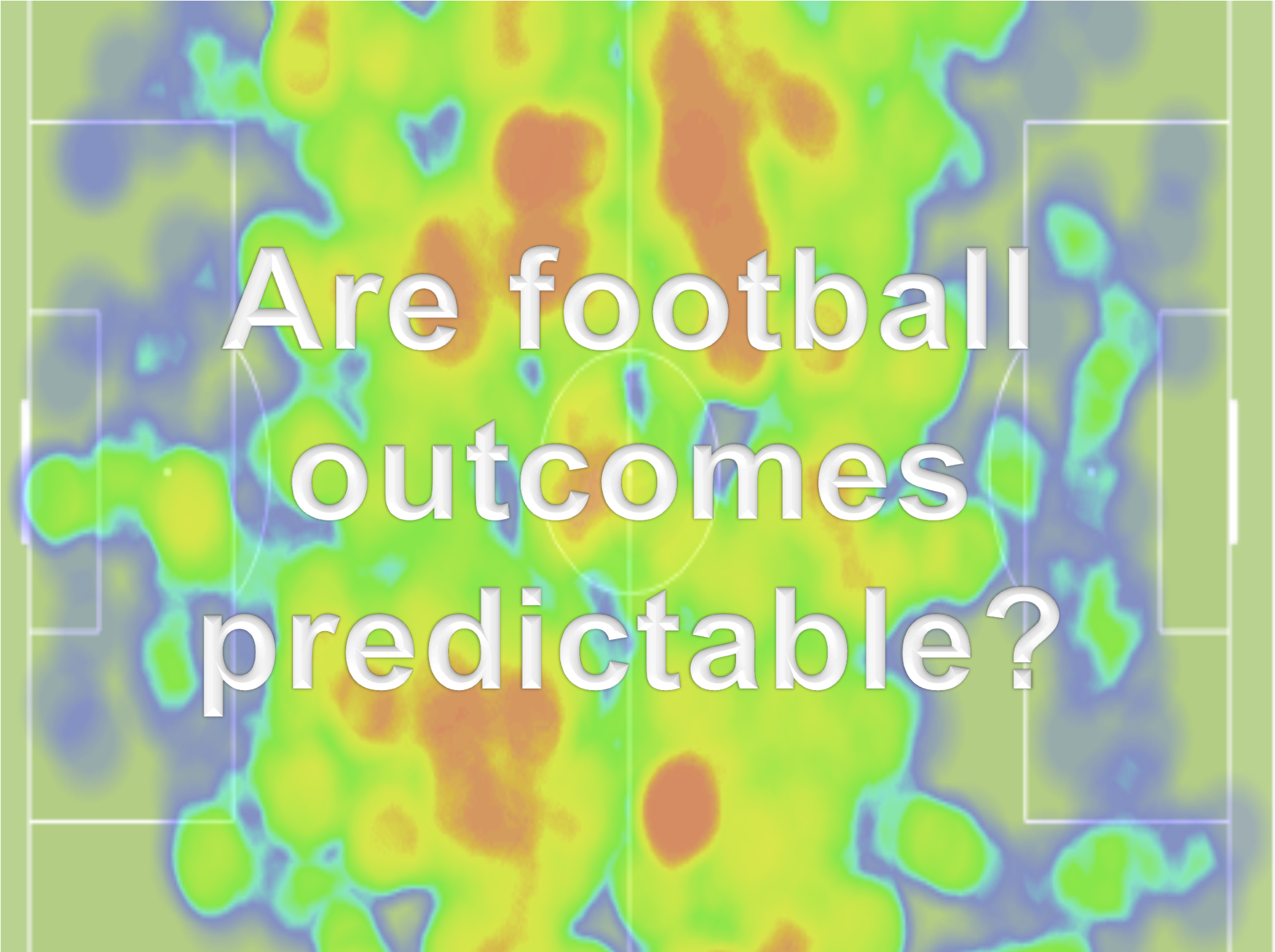
Networks & Sport Data Mining

[Cintia, Pappalardo et al at KDD LAB]

The patterns of success in football

“Football is a simple game: 22 men chase a ball for 90 minutes and at the end, the Germans always win” -- Gary Lineker (after Italy 1990 Final)

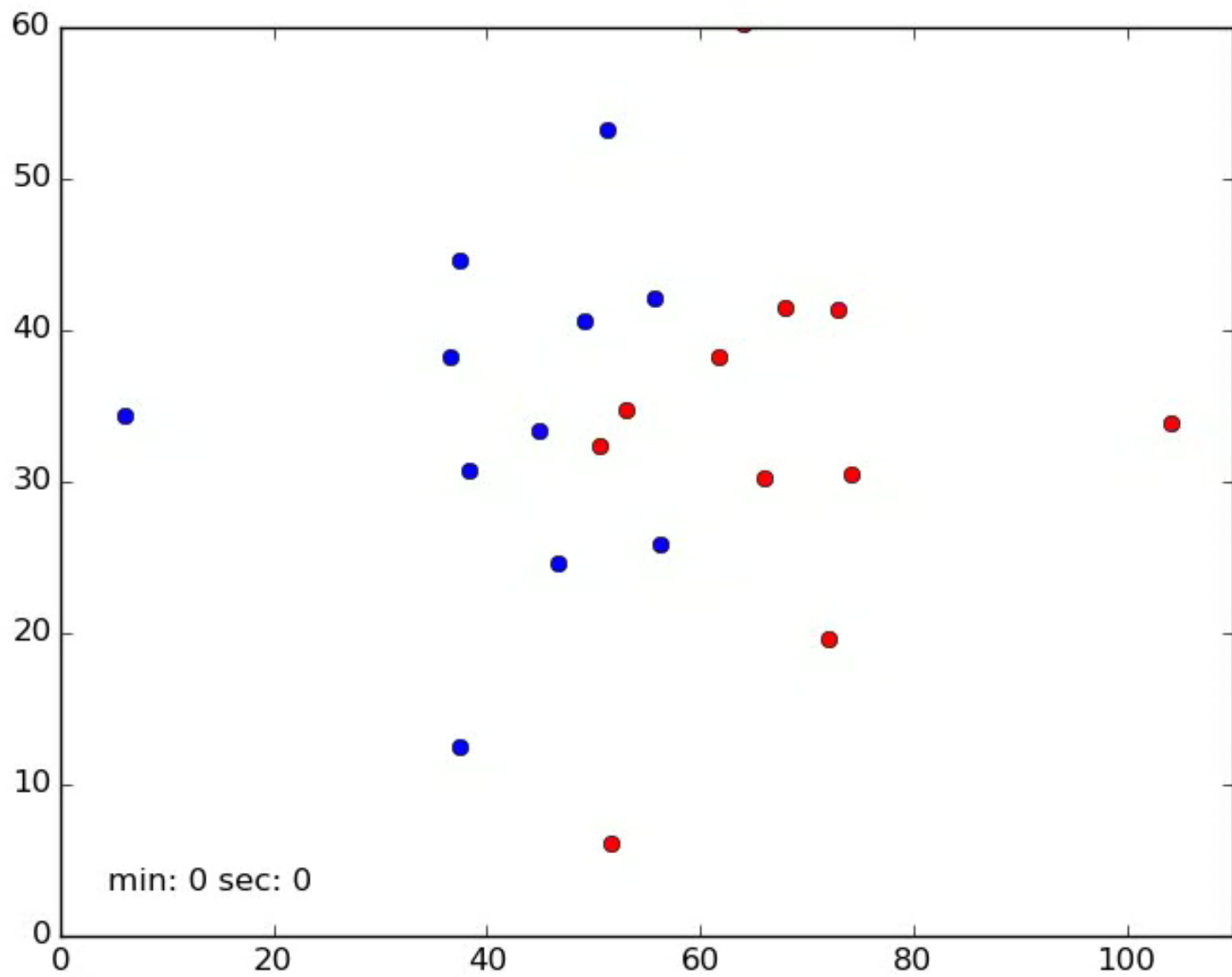


A heatmap visualization of a football pitch, where colors represent different levels of activity or probability. The pitch is outlined with white lines, including the center circle and the three main rectangular zones (defence, midfield, and attack). The color scale ranges from blue (low activity) to red (high activity). High-activity areas (red/orange) are concentrated in the central midfield and the attacking half, particularly around the center circle and the penalty area. The defensive half shows lower activity (green/blue).

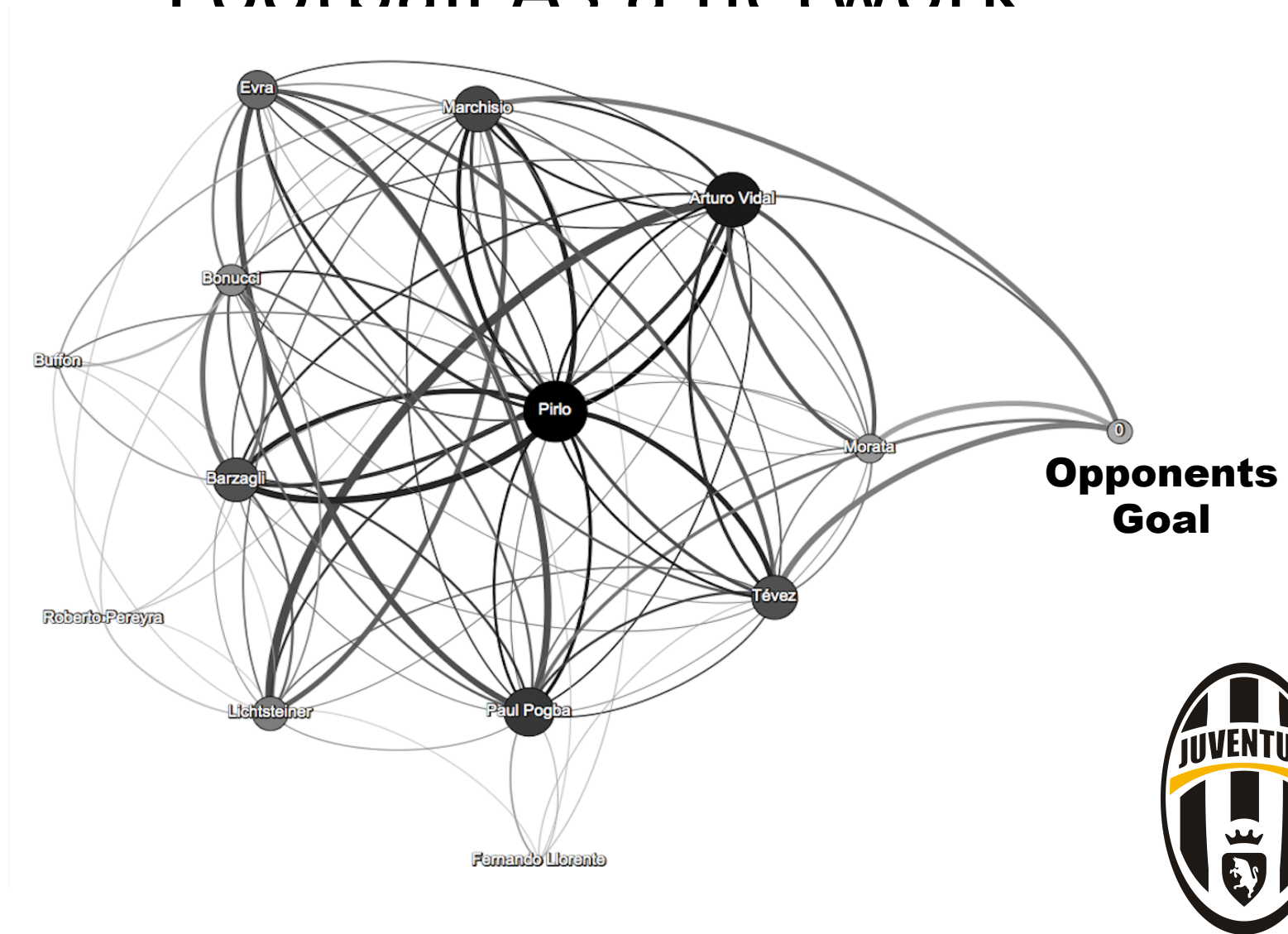
Are football
outcomes
predictable?

Data from each single match

```
...  
<tackle,15.4,41.1,112>  
<pass,25.0,67.1,113>  
<pass,65.0,87.1,115>  
<assist,82.1,35.8,120>  
<goal attempt,82.1,35.8,121>  
.....
```

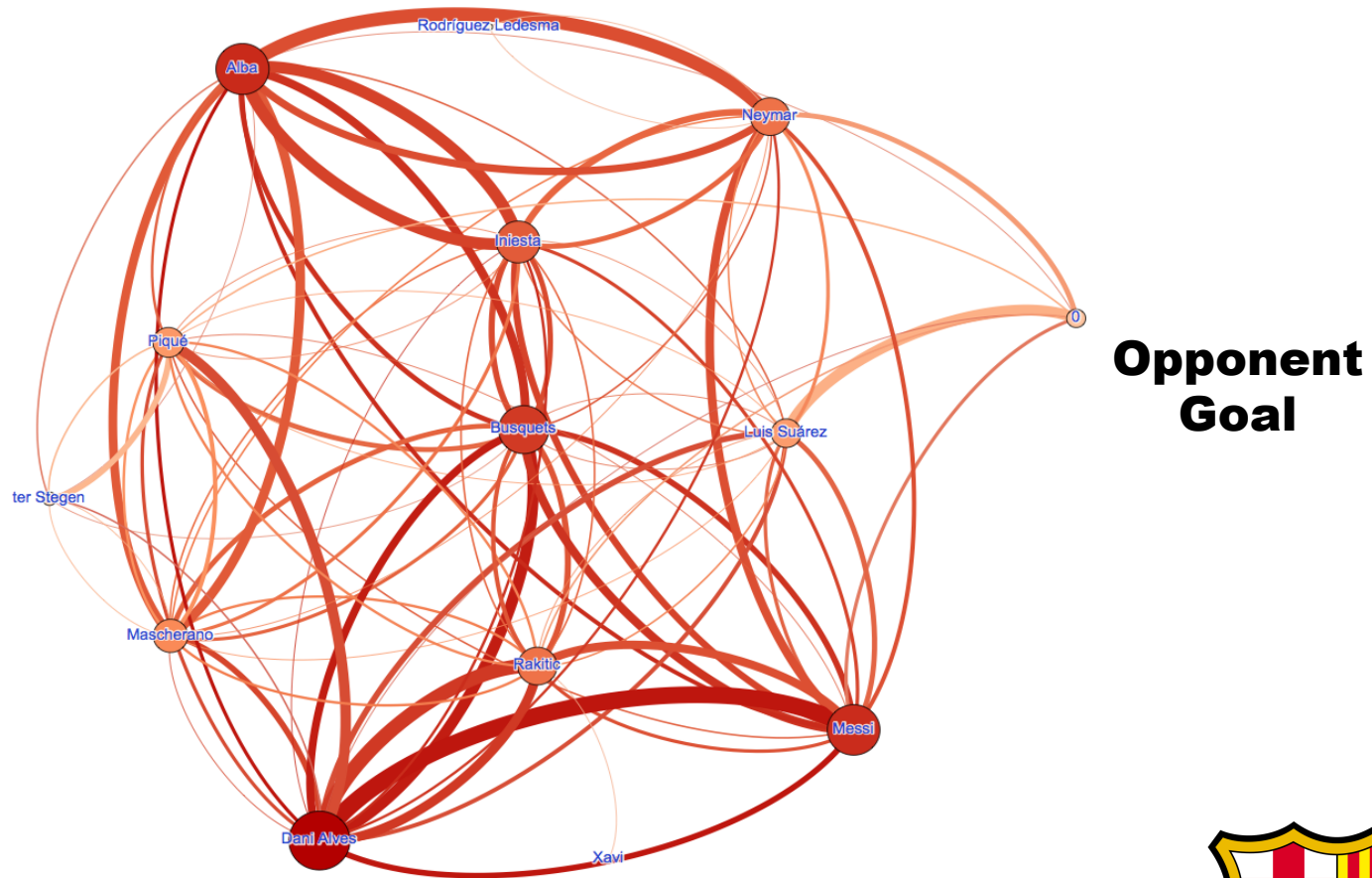


Football As a network



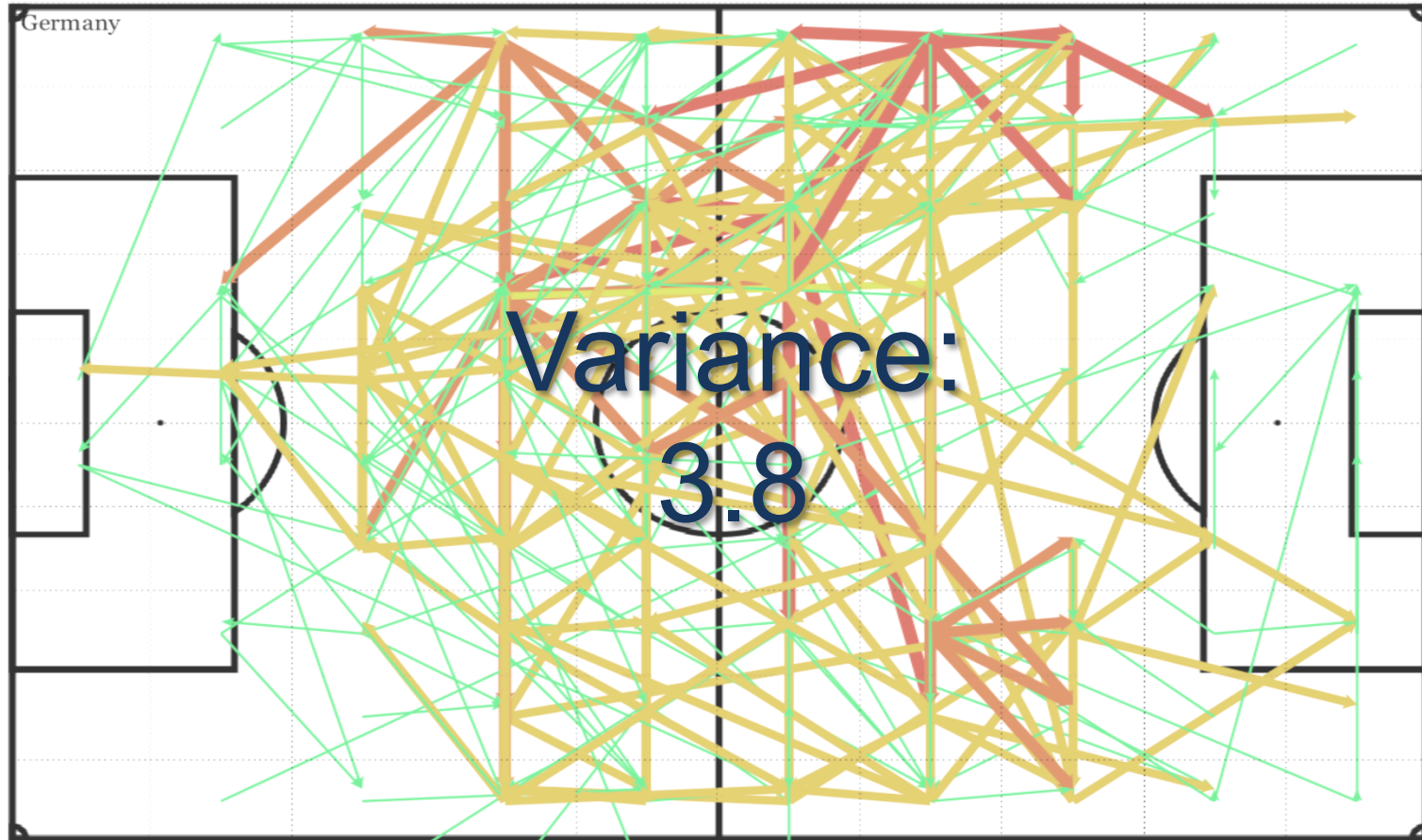
Juventus passes network from last champions league final

Football as a Network



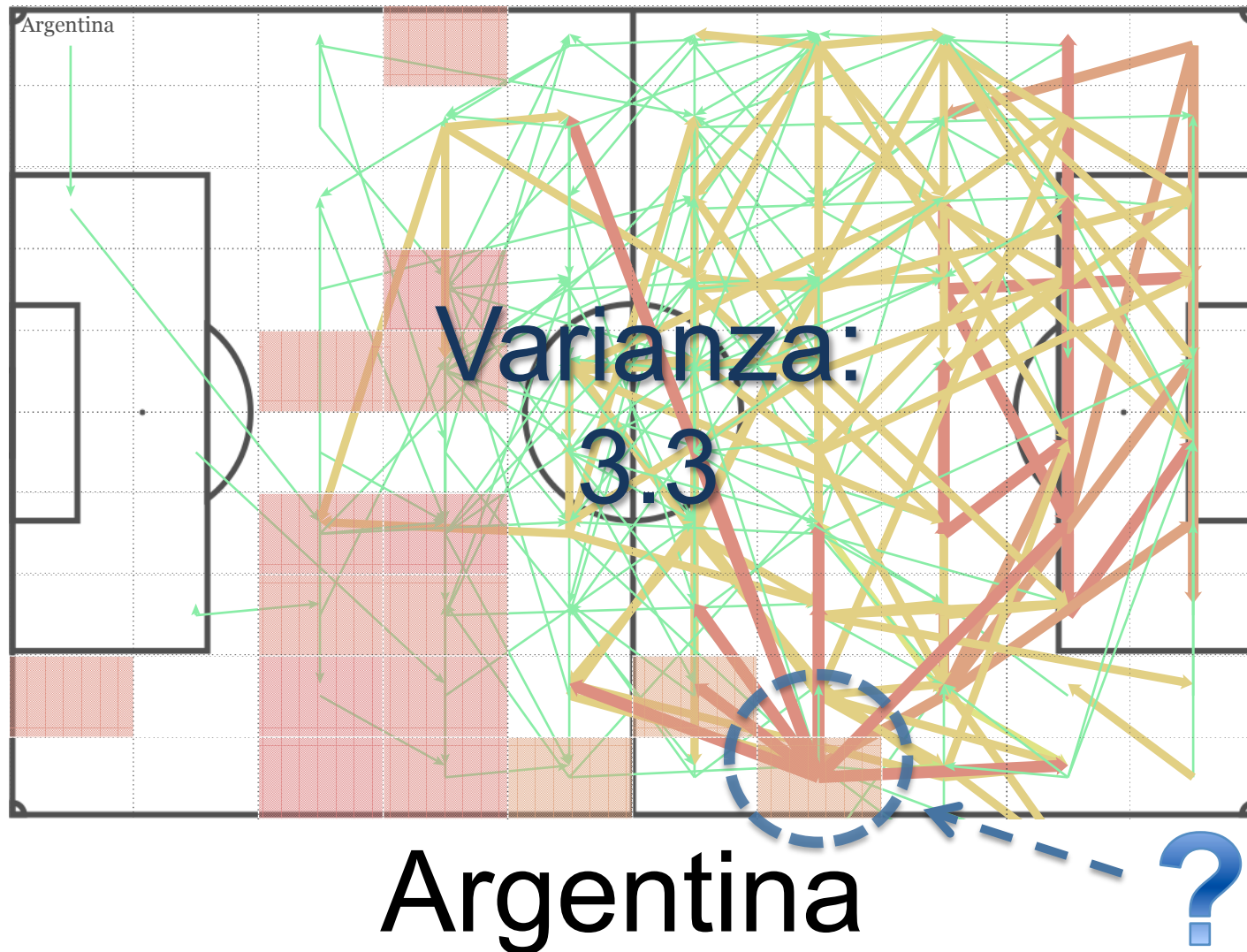
Barcelona passes network from last champions league final

The passes network among zones



Germany

The passes network among zones

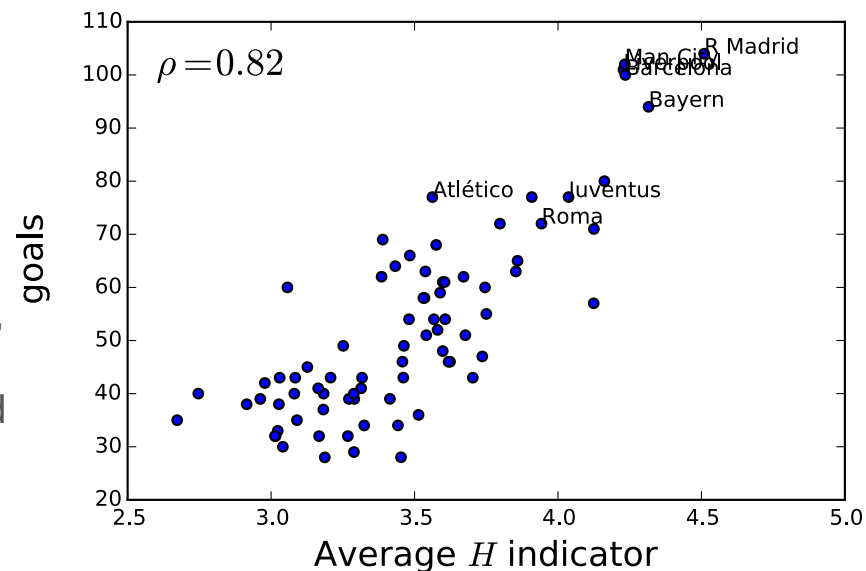


Network analysis for performance evaluation

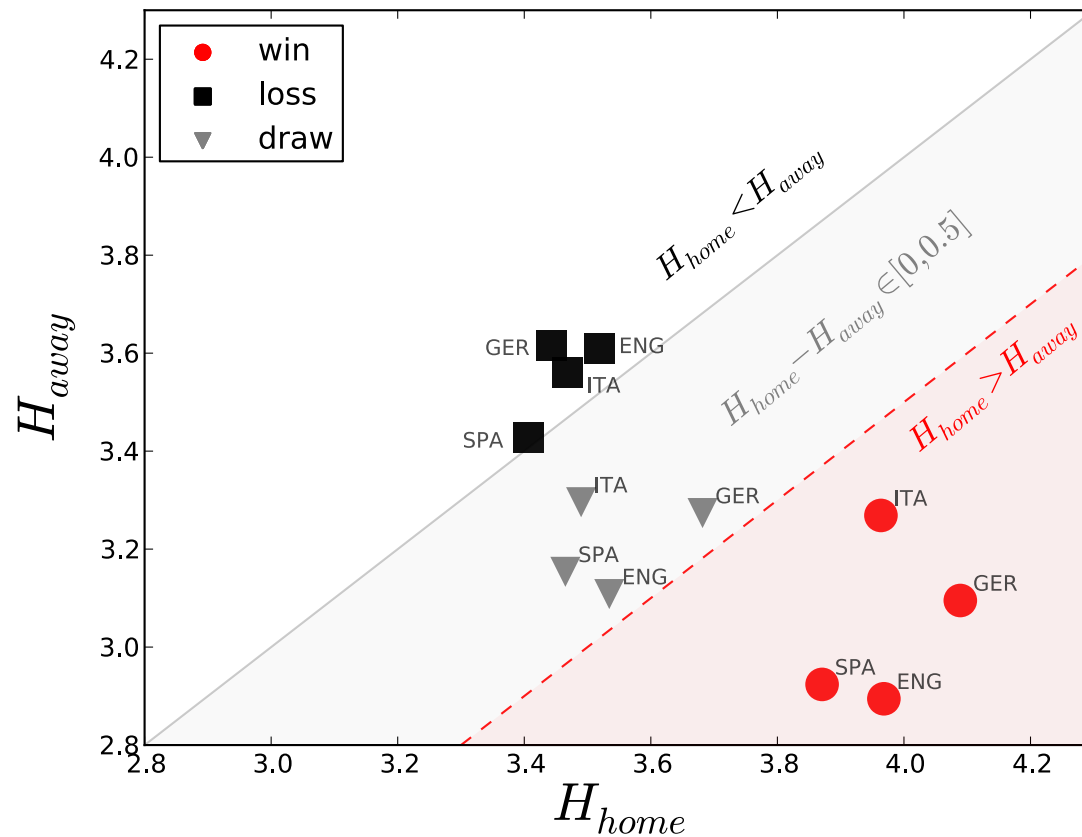
- Networks characteristics as a proxy for performance evaluation and outcome prediction

measure	description
w	total passing volume
μ_p	mean players' passing volume
σ_p	variance of players' passing volume
μ_z	mean zones' passing volume
σ_z	variance of zones' passing volume
H	combination of above measures

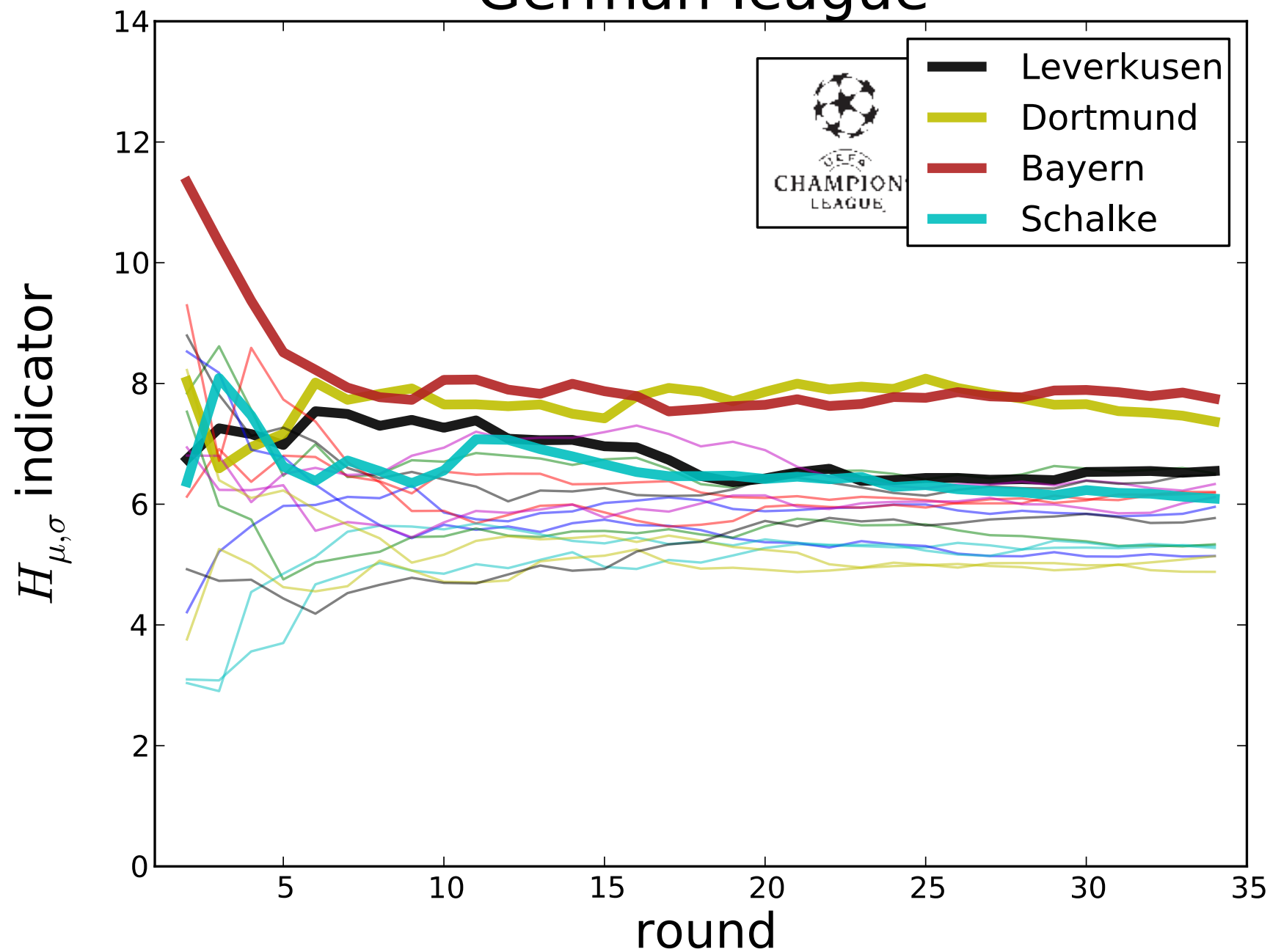
The harsh rule of the goals”: Big Data analytics and football team success
Paolo Cintia, Luca Pappalardo, Dino Pedreschi, Fosca Giannotti and Marco Malvaldi
2015 IEEE/ACM Int. Conf. on Data Science and Advanced Analytics (DSAA'2015)



Predicting game outcomes



German league





simulated ranking		real ranking	
Bayern	91	Bayern	90
Leverkusen	72	Dortmund	71
Dortmund	68	Schalke	64
Wolfsburg	59	Leverkusen	61
Augsburg	58	Wolfsburg	60
Hoffenheim	49	Mönchengladbach	55
Hertha	49	Mainz	53
Mainz	48	Augsburg	52
Schalke	47	Hoffenheim	44
Frankfurt	46	Hannover	42
Mönchengladbach	42	Hertha	41
Hannover	41	Werder	39
Hamburg	38	Freiburg	36
Stuttgart	35	Frankfurt	36
Freiburg	31	Stuttgart	32
Werder	24	Hamburg	27
Braunschweig	22	Nürnberg	26
Nürnberg	17	Braunschweig	25



CNR

**La dura
legge dei DATI!**



Consiglio Nazionale
delle Ricerche

SobigData
Euro Lab on Big Data Analytics
& Social Mining

**Paolo Cintia
Marco Malvaldi
Luca Pappalardo**

con la partecipazione di
**Dino Pedreschi
Fosca Giannotti
Salvatore Rinzivillo**

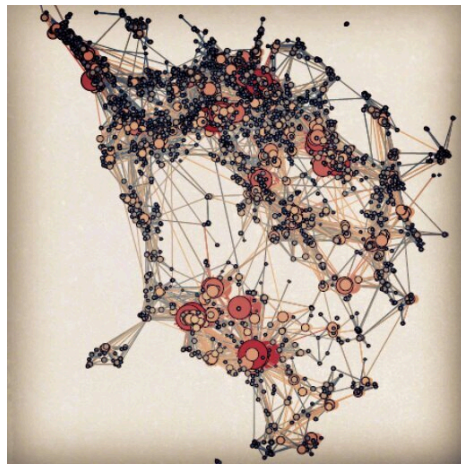
...CONTINUA?

We are the champions – understanding the patterns of success in football

P.Cintia, L. Pappalardo, F. Simini

Kdd Lab ISTI CNR Pisa – University of Pisa – University of Bristol

www.kddlab.isti.cnr.it



Big Data Tales

Adventures in data

www.bigdatales.com



@bigdatatales

Analysing and mining the retail market as a complex system

Michele Coscia, Fosca Giannotti, Mirco Nanni, Dino Pedreschi, Diego Pennacchioli, Salvatore Rinzivillo

KDDLab ISTI-CNR & University of Pisa
IMT Lucca, CID Harvard University



Supermarket transaction data

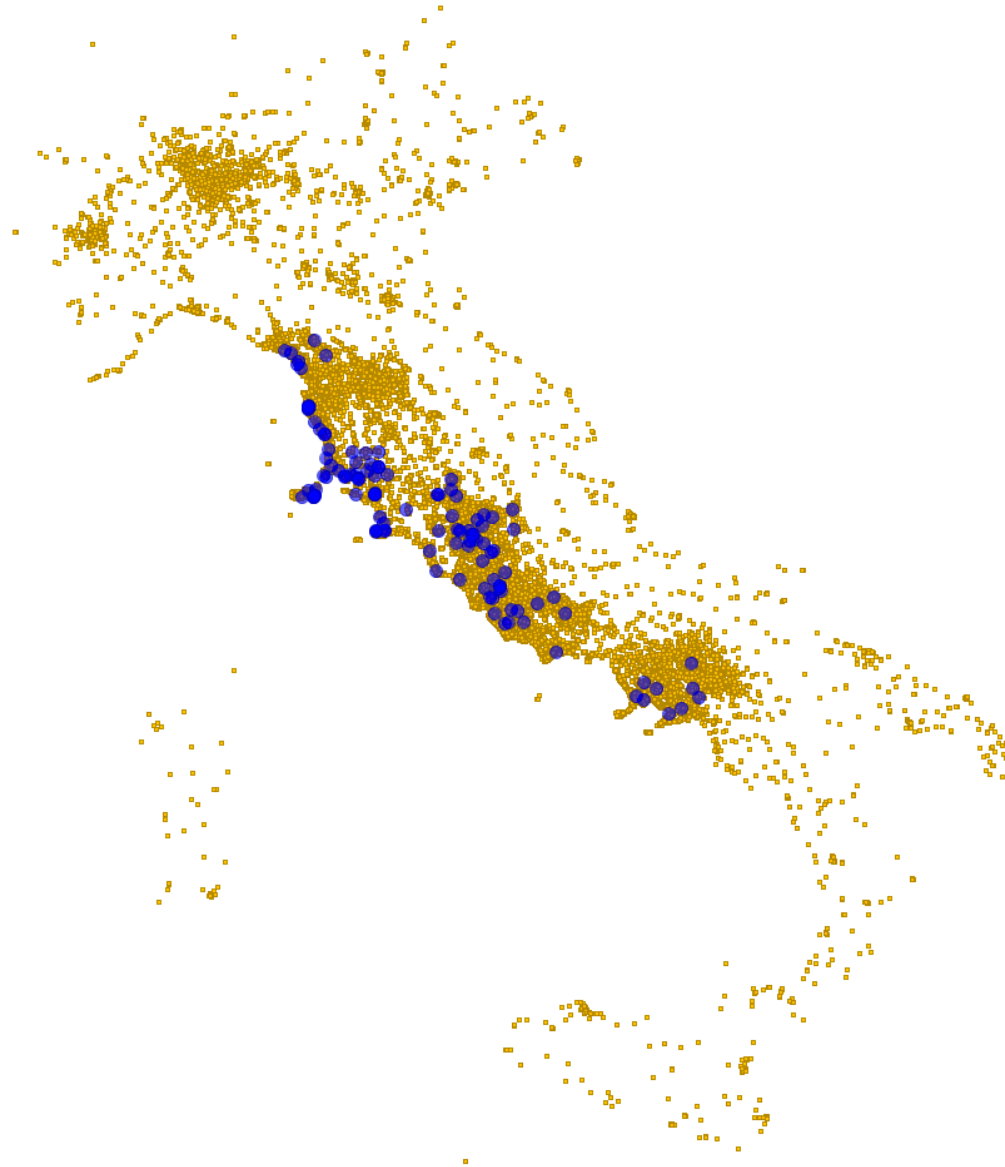
- 1,066,020 customers
- 345,208 products
- 138 stores in 4 Italian regions
- Since January 1st, 2007

MISSION (extract):

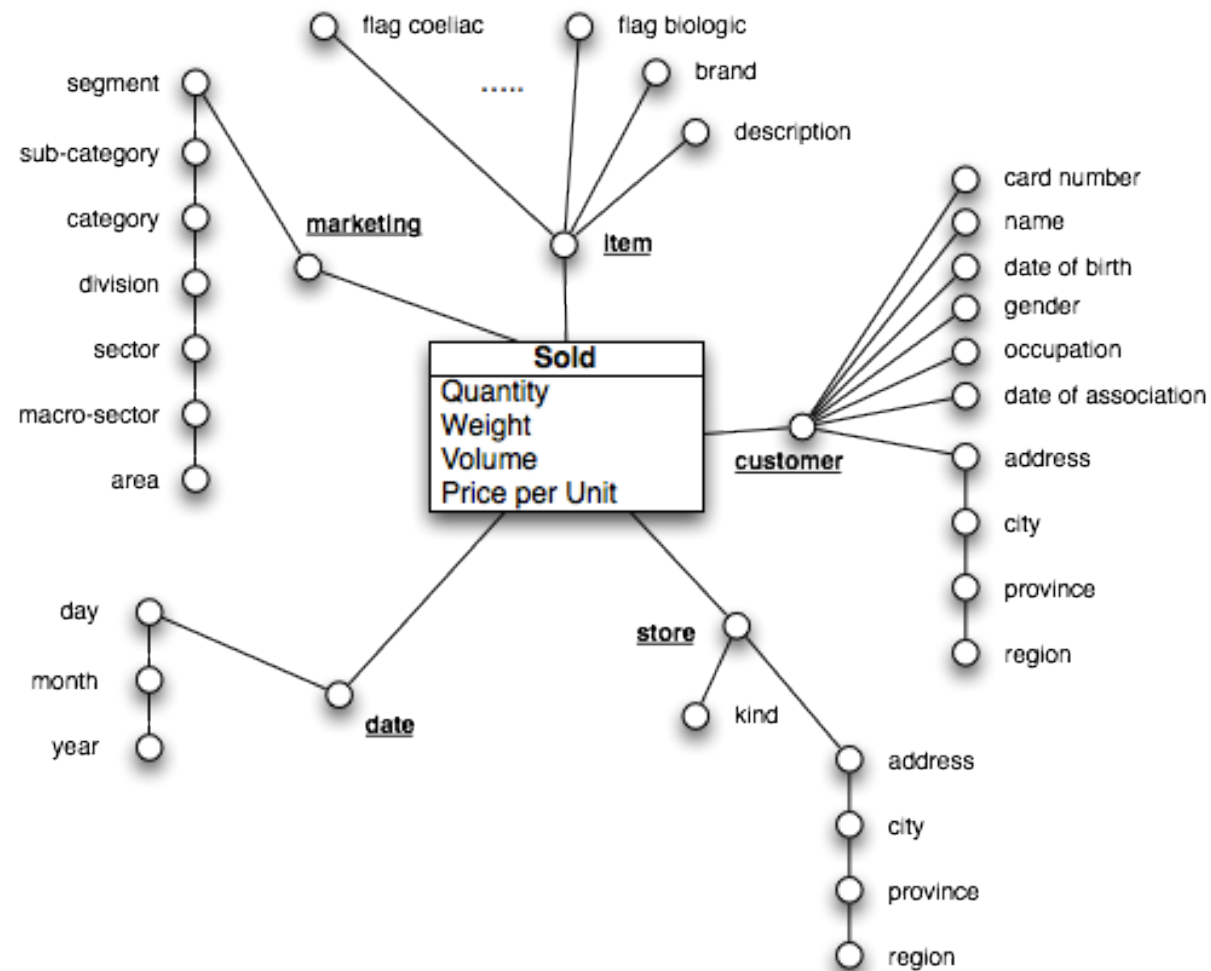
- defend the family budget, offering quality goods and services at the best condition
- promote the spirit of solidarity among customers, employees and their families, offering for their free time socially useful initiatives
- protect health and safety of consumers
- help developing Nations and weaker social categories. even through Fair



Customer/shop Geographical Distribution



Data Model



7,003 marketing segments

The retail market as an ecosystem

Coscia, Pennacchioli, Giannotti,
Pedreschi, Rinzivillo
EPJ Data Science 2014

Data Selection

- Data from 2007-2009 (period 2009-2011 used as control)
- Only the customers from Livorno
- We aggregate products to their marketing Segment
- We clean not sold or meaningless products
- We fill a 317,269 customers X 4,817 products with the amount purchased (182,821,943)

Census (2011)	343K
registered	212K
At least once	146K
Actives 07-11	102K
Actives 2011	77K

$343K / 77K = 4.45$
(2.5 persons x family in AVG) [ISTAT]

Data Preparation

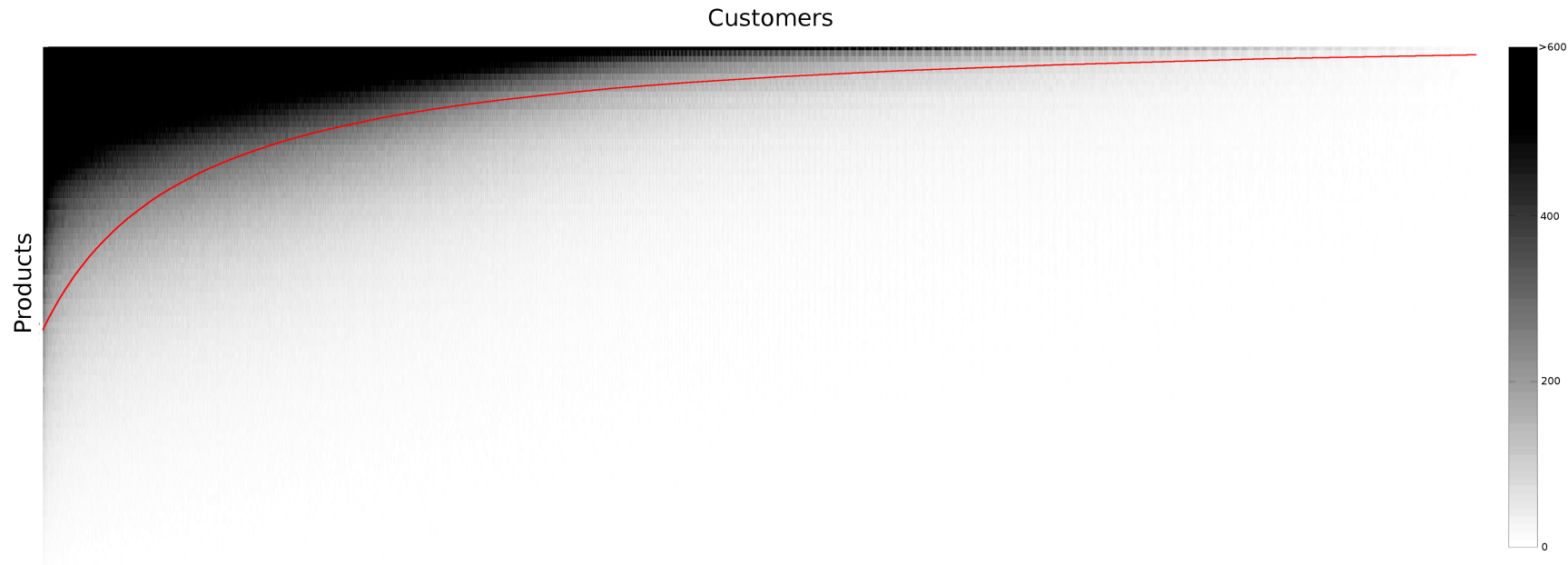
- Sort rows and columns according to their sum

$$\text{lift}(X, Y) = \frac{\text{supp}(X, Y)}{\text{supp}(Y) \times \text{supp}(X)}$$

$$M_{cp} = \begin{cases} 1 & \text{if } \text{lift}(c_j, p_i) > 1; \\ 0 & \text{otherwise.} \end{cases}$$

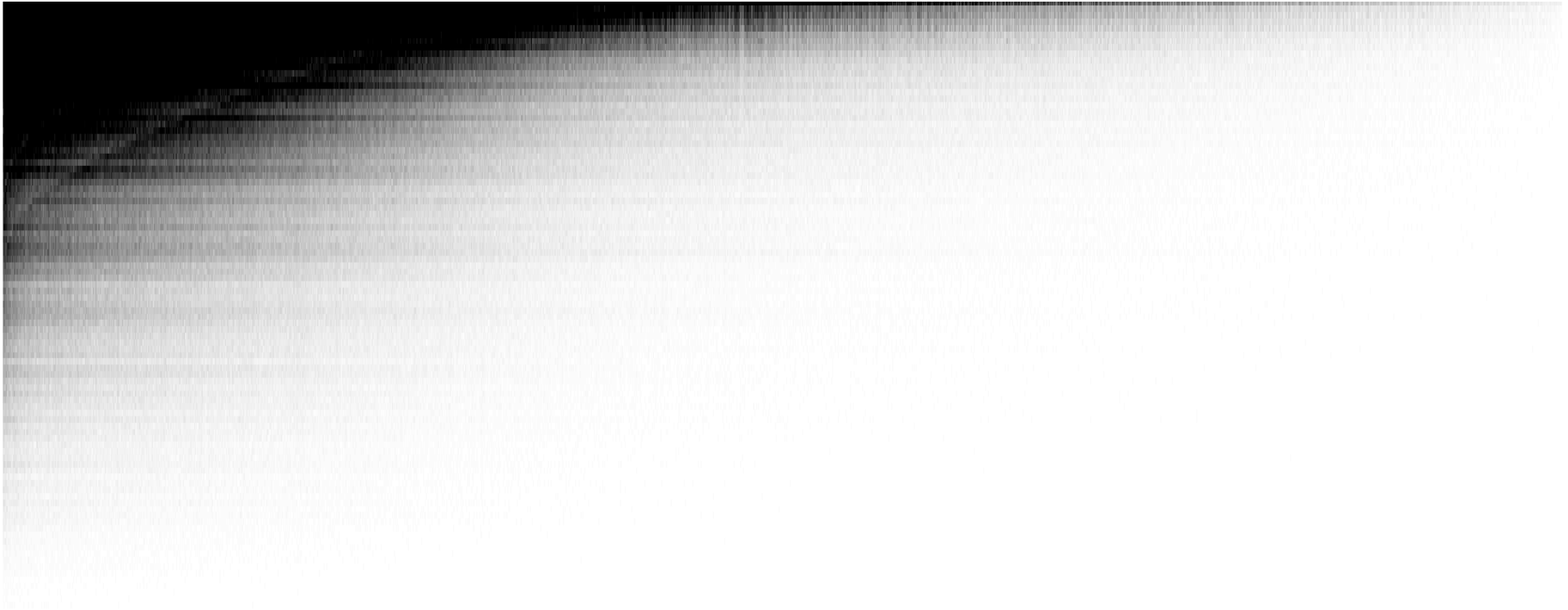
- 37,338,591 ones (2.4431697% fill)

The Retail Purchase Matrix



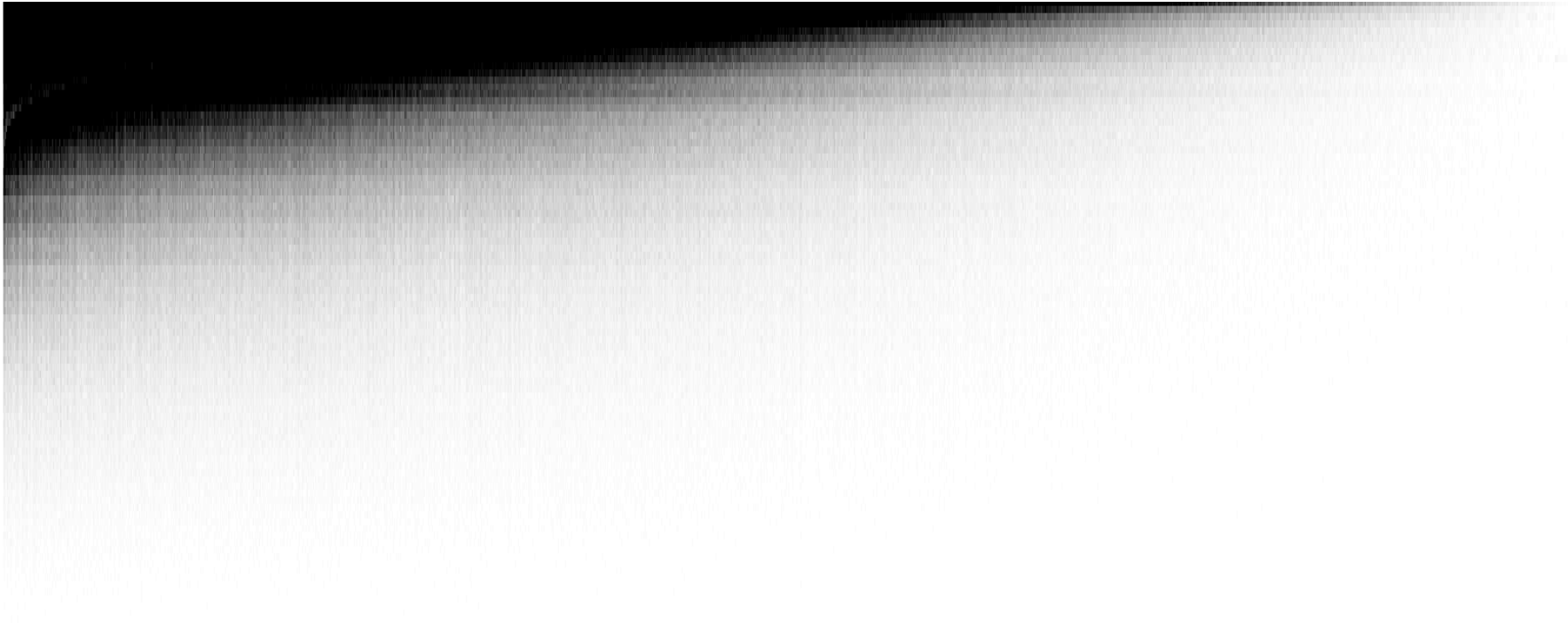
- There are products bought by everyone
- There are customers buying everything
- The volume of sales of a product is identifying who is buying it
- ... is this phenomenon robust?

Total Price instead of Amount of Items

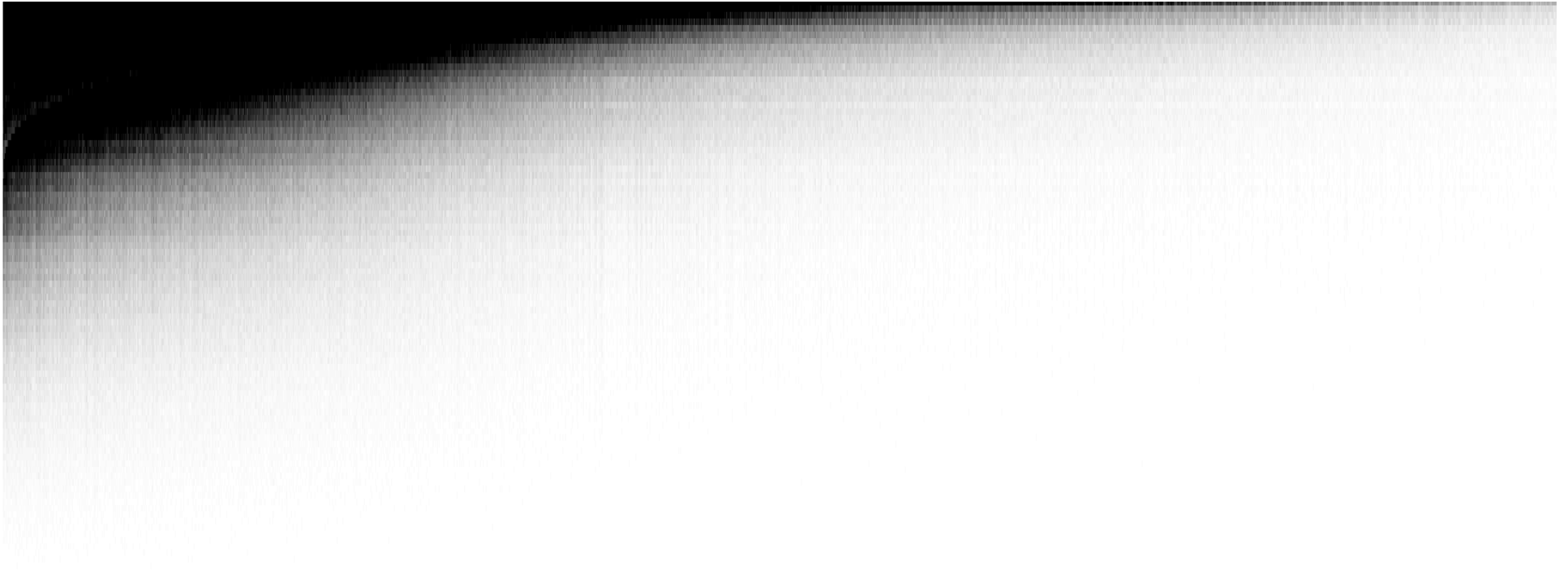


Category instead of Segment

Lazio matrix



Livorno 2009-2011



... is this phenomenon expected?

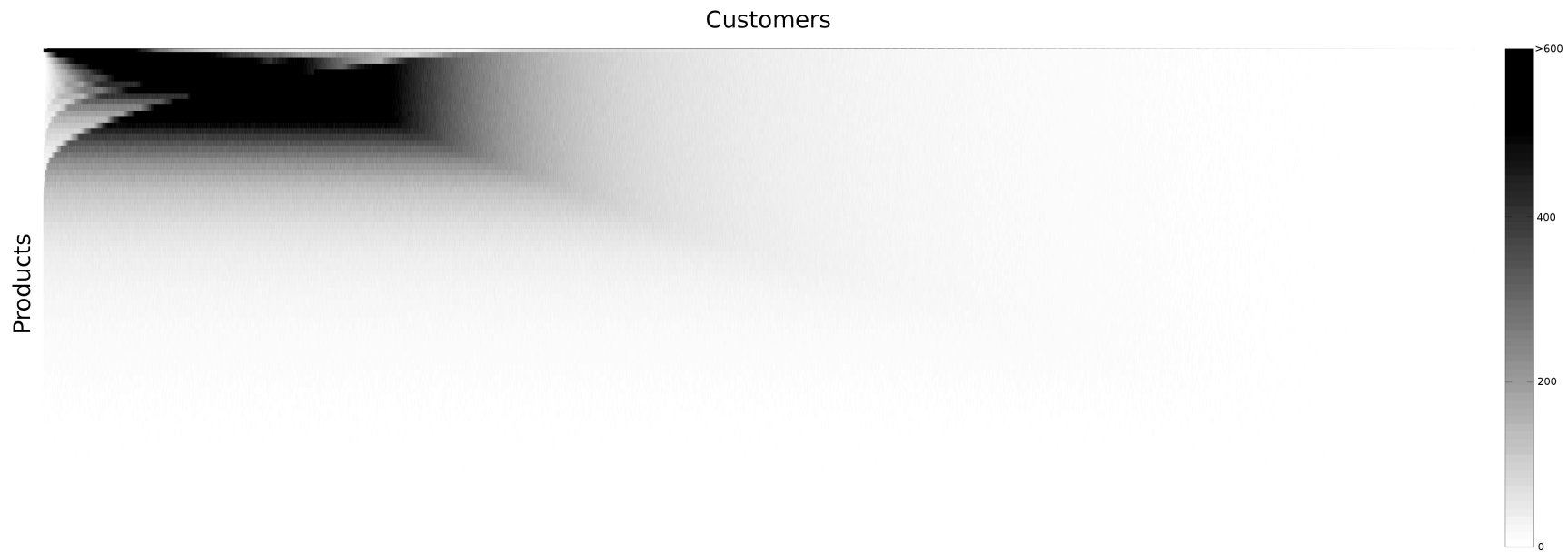
Null Hypothesis

Given the volume of sales of products, and the quantity of products bought by customers, the nestedness of the matrix is inevitable even in a random world where there is no relation between products and customers

Null Model

- The purchases are distributed randomly;
 - Customers preserve the total amount of their purchases;
 - Each product preserves its sale volume
-
- We randomize the matrix by preserving the sums of columns and rows

Null Model Matrix



Retail Purchase Law

Average of the one-density at the left of the isocline and zero-density at the right of the isocline

$$N(M_{c,p}, f_*) = \frac{1}{2} \left(\frac{f_l(M_{c,p}, 1)}{f_l(M_{c,p}, *)} + \frac{f_r(M_{c,p}, 0)}{f_r(M_{c,p}, *)} \right)$$

f_*	$N(M_{c,p}, f_*)$
$ax + b$	0.616106811666
$ax^2 + bx + c$	0.628533747603
$a \log(x) + b$	0.623911138356
ax^b	0.572996769269
$\frac{a}{x}$	0.609588181022
$-\frac{ax+b}{cx+d}$	0.632547410976

$$\alpha j + \beta i + \gamma ij + \delta = 0$$

$$i = -\frac{\alpha j + \delta}{\gamma j + \beta}, \quad j = -\frac{\beta i + \delta}{\gamma i + \alpha}$$

$$\alpha = 2342.5757, \quad \beta = 14.4349, \quad \gamma = 0.0720, \quad \delta = -5535548$$

Product Relationships

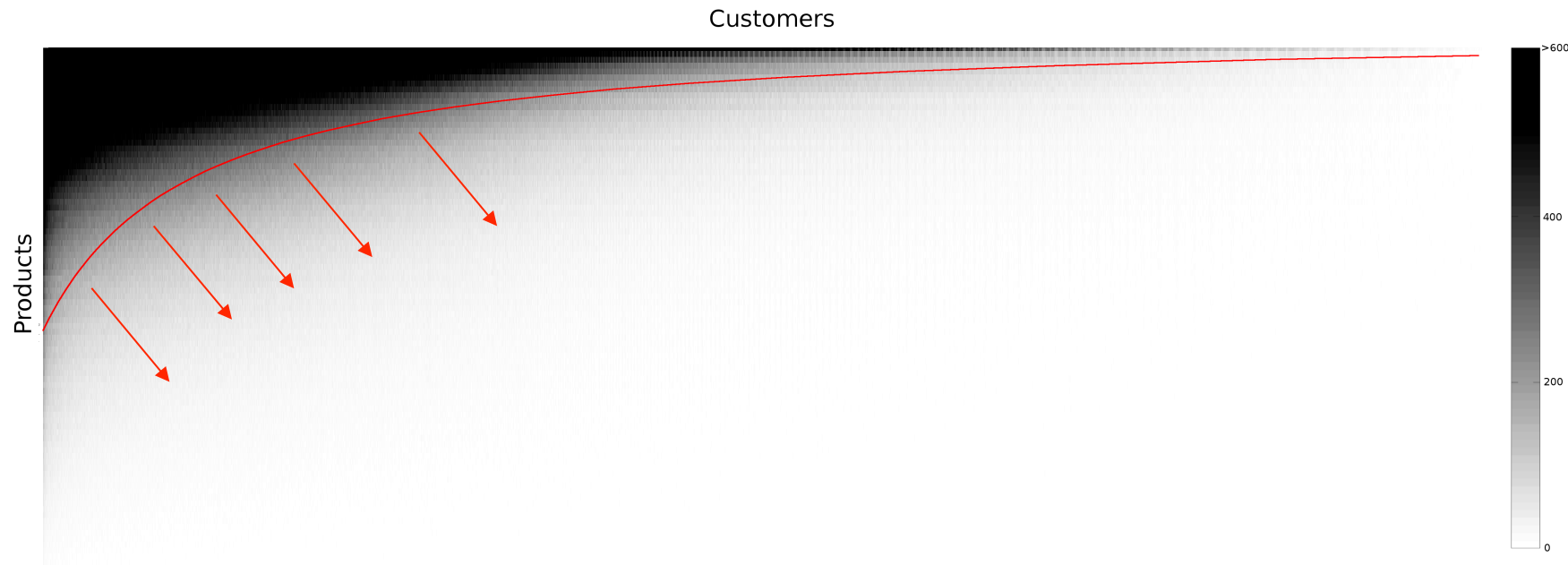
- Before: $A \Rightarrow B$, association rule
- A and B co-appears unexpectedly X times, therefore they are related
- Now: we introduce an ecology approach
- If product B is directly dependent of A's ecosystem, then A and B are related

Product Relationships

p_i	p_{i-1}	$P(p_i)$	$P(p_i p_{i-1})$
Dishwasher Salt	Dishwasher Soap	8.39%	30.41%
Asparagus	Olive	8.00%	26.12%
Peppers	Chicory	7.31%	23.73%
Canned Soup	Preserved Anchovies	9.96%	32.23%
Wafers	Sugar Candies	11.30%	21.67%

$$\frac{P(p_i|p_{i-1})}{P(p_i)} = 1.993 \text{ in avg}$$

Identifying the “willing customer” set



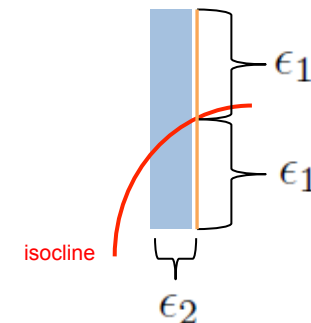
- We select the customers on the isocline: how many of them are willing to buy p ?
- How many random customers we need to select without the isocline to have a set of the same size?

Product Relationships

p_i	$ TC^* $	$ TC $	$\frac{ TC_r }{ TC }$
Tomino Cheese	58	137	7.51095
Raw Ham	78	144	5.81250
Apricot Jam	66	127	4.66142
Anchovies	83	144	4.06250

$$TC \quad j - \epsilon_1 \leq j' \leq j + \epsilon_1 \quad \text{and} \quad M_{cp}(c_j, p_i) \neq 1$$

$$TC^* \quad \exists x, M_{cp}(tc, p_x) = 1 \quad \text{with} \quad i - \epsilon_2 \leq x < i$$



by fixing $\epsilon_1 = 100$ and $\epsilon_2 = 2$ averages of the $\frac{|TC_r|}{|TC|}$ ratio is 3.55594

The hierarchy of Needs

- What are the most basic products?
- What are the most sophisticated products?
- Not only biggest volume sales or higher prize
- To be basic, p should be bought by the least “complex” customers
- To be sophisticated, p should be bought by the customers with more sophisticate need

The product sophistication

- we calculate the sums of the purchase matrix for each product and customer
- we need to **correct these sums recursively**: we need to calculate the avg level of sophistication of the **customers'** needs by looking at the avg sophistication of the **products** that they buy, and then use it to **update** the avg sophistication of these products
- so, we take the eigenvector associated with the second largest eigenvalue (that is associated with the variance in the system) [HITS variation]

Product Sophistication

Correct the ubiquity
of a product

$$k_{N,p} = \frac{1}{k_{0,p}} \sum_c M_{cp} k_{c,N-1}$$

With the complexity
of customers buying it

$$k_{N,p} = \frac{1}{k_{0,p}} \sum_c M_{cp} \frac{1}{k_{c,0}} \sum_{p'} M_{cp'} k_{N-2,p'}$$

That needs to be
recursively corrected

$$k_{N,p} = \sum_{p'} k_{N-2,p'} \sum_c \frac{M_{cp} M_{cp'}}{k_{0,p} k_{c,0}}$$

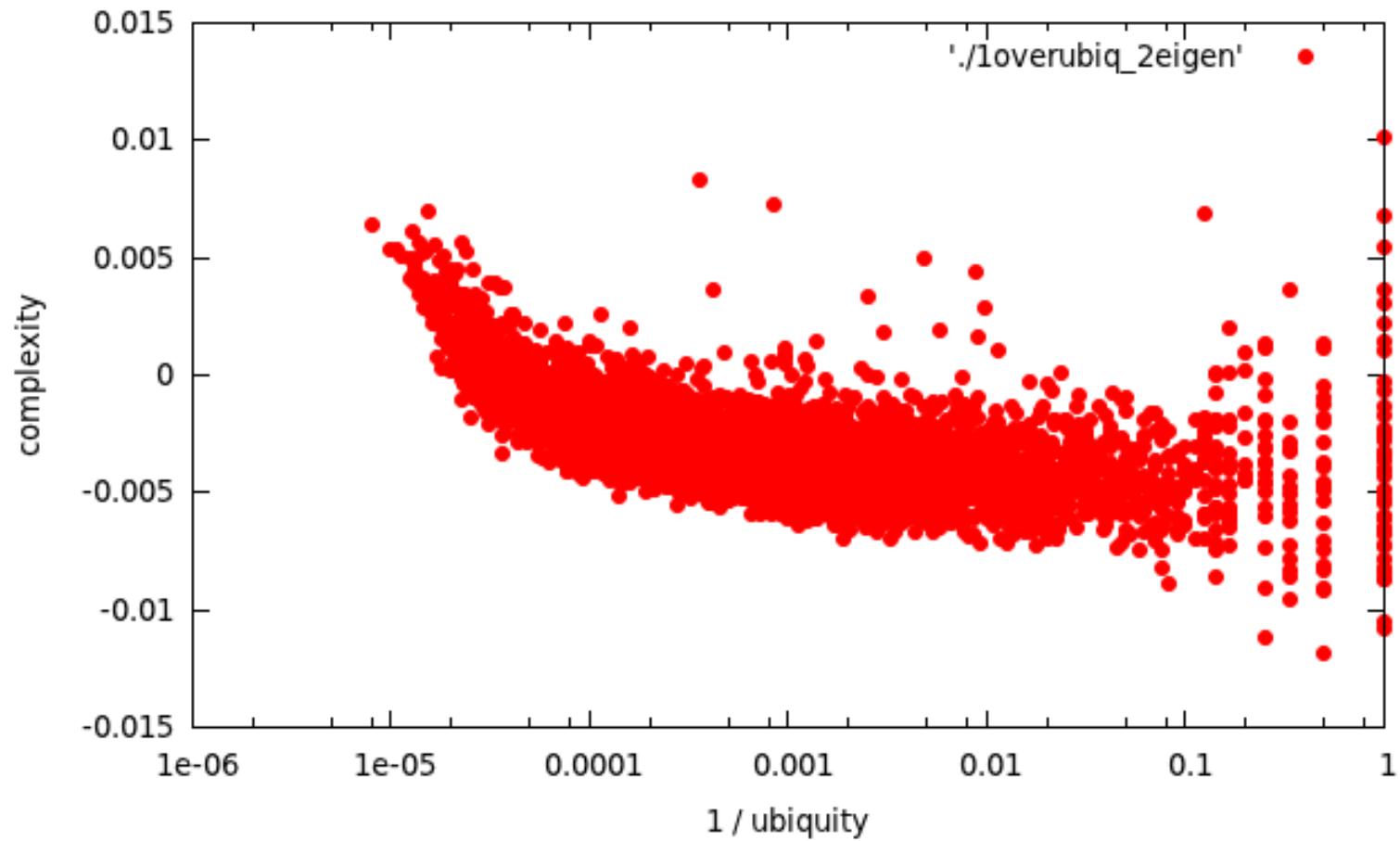
With the ubiquity of the products
they buy

$$k_{N,p} = \sum_{p'} \widetilde{M}_{pp'} k_{N-2,p'}$$

It is equivalent to the eigenvector
Of the corrected matrix

$$\widetilde{M}_{pp'} = \sum_c \frac{M_{cp} M_{cp'}}{k_{0,p} k_{c,0}}$$

PS vs Volume Sale



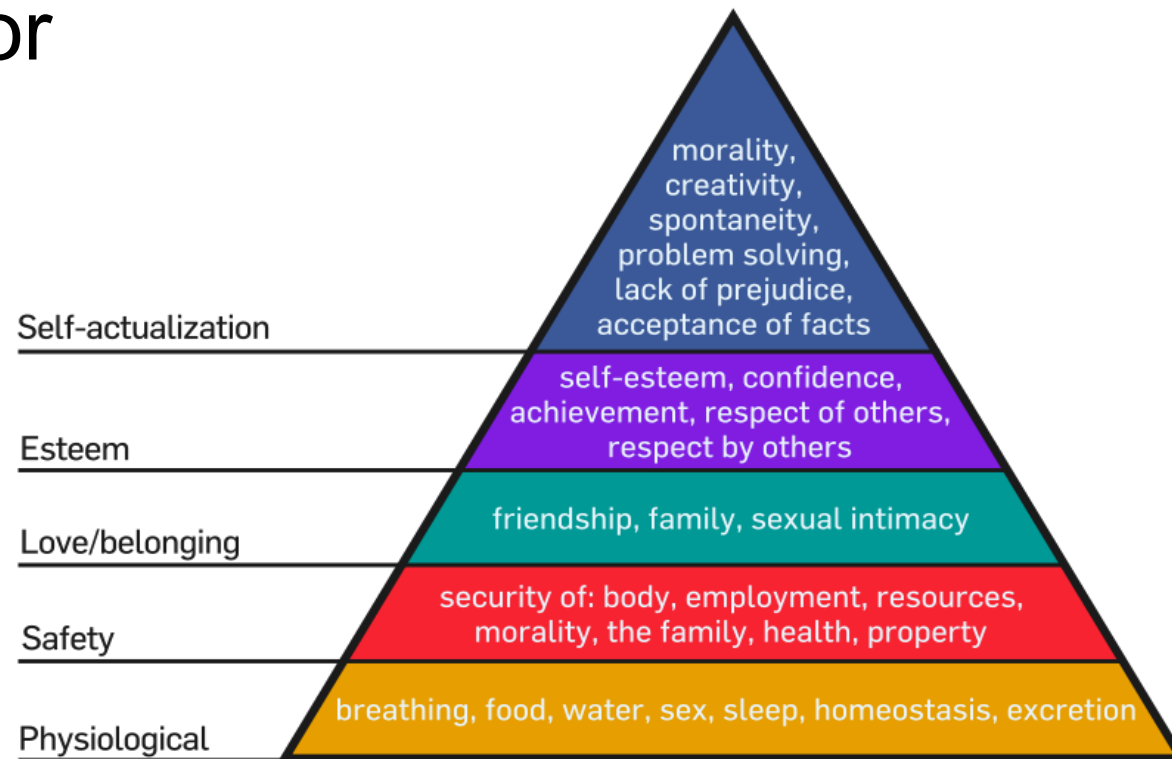
Least and Most Sophisticated Products

p_i	PS
Regular Bread	-4.41
Natural Still Water	-4.19
Yellow Nectarines (Peaches)	-3.84
Semi-Skimmed Fresh Milk	-3.81
Bananas	-3.53

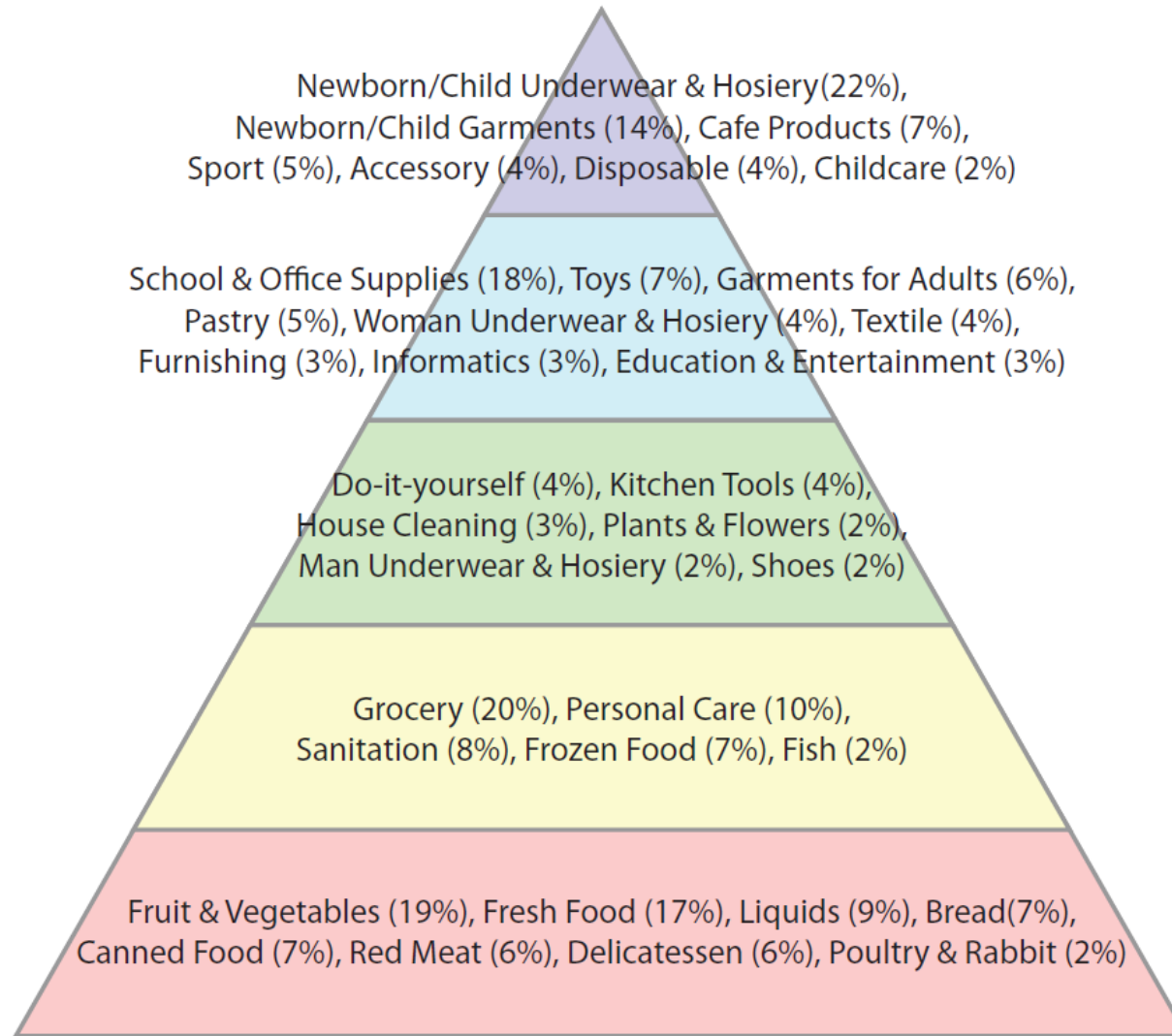
p_i	PS
LCD 28"/30" Televisions	2.91
DVD Music Compilations	2.86
Sauna clothing	2.66
Jewelry Bracelets	2.53
RAM Memories	2.33

Building the Pyramid

- Introduced by Abraham Maslow in 1954
- Monodimensional segmentation of the eigenvector
- K-means
- $K = 5$



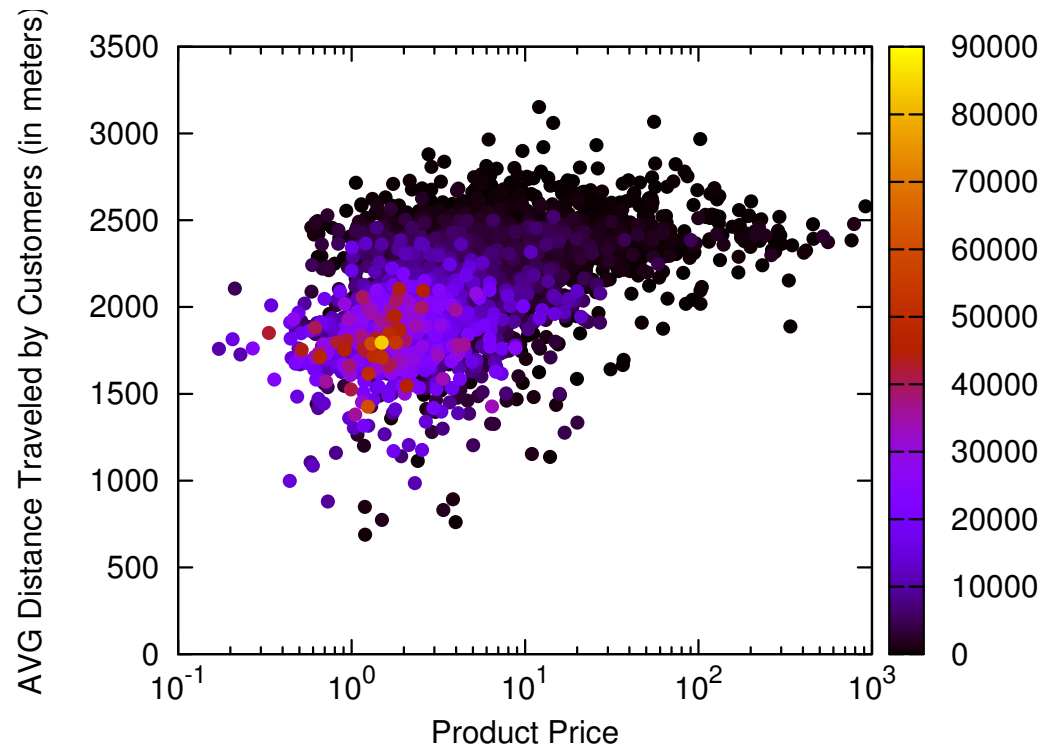
The Hierarchy of Needs



Applicazione spaziale [1]

- each dot is a purchase representative
- if a customer bought products of the same price in different shops, than the distance is weighted with the price

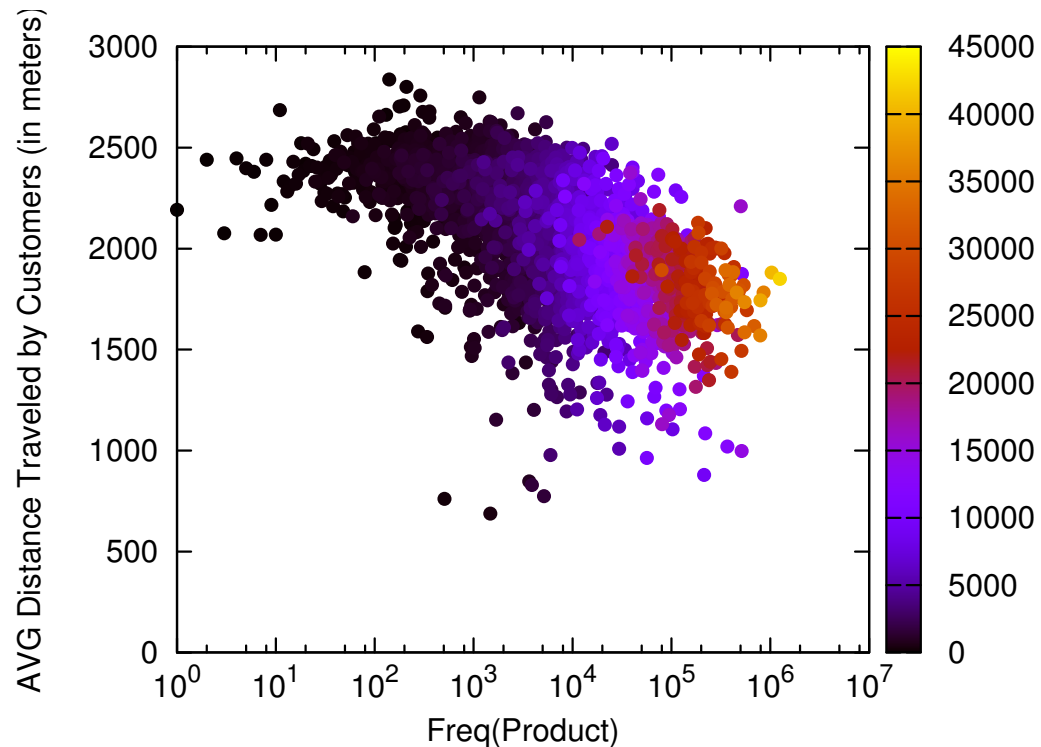
$$d(c_i, p_j) = \sum_{\forall s \in S} \frac{p_j(c_i, s) \times d(c_i, s)}{p_j(c_i, *)}$$



log-normal regression $f(x) = a \log x + b$ $R^2 = 17.25\%$

Applicazione spaziale [1]

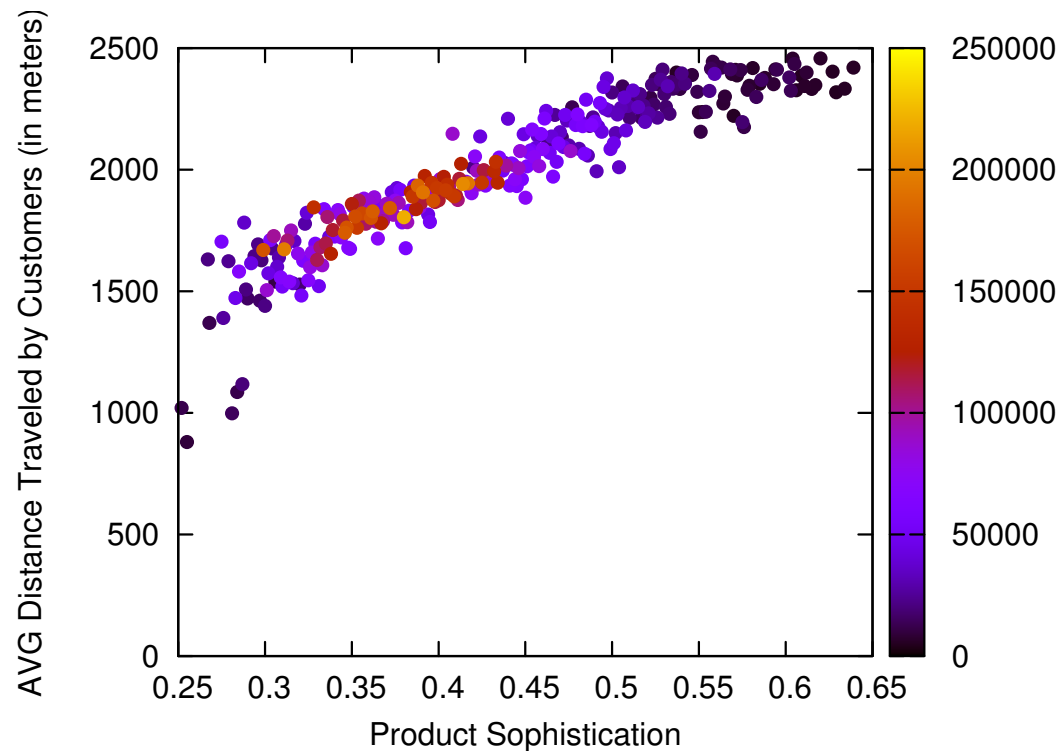
- each dot is a purchase representative
- if a customer bought products of the same frequency in different shops, than the distance is weighted with the frequency



log-normal regression $f(x) = a \log x + b$ $R^2 = 32.38\%$

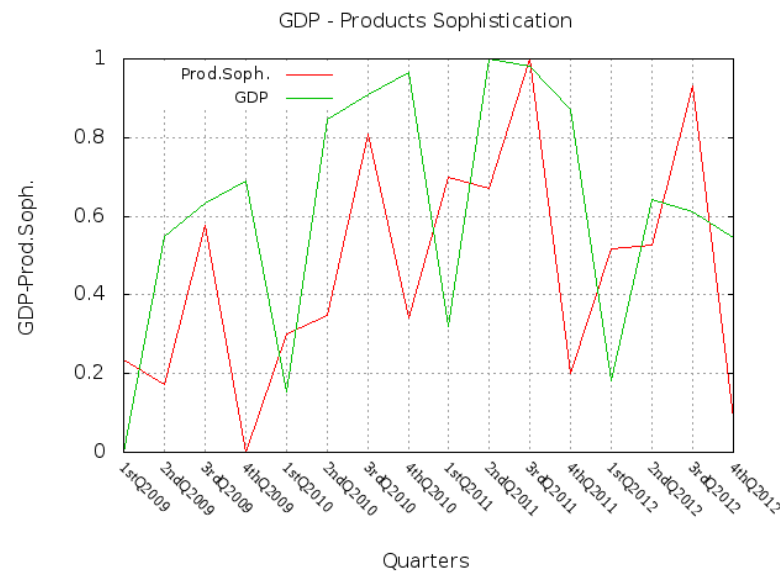
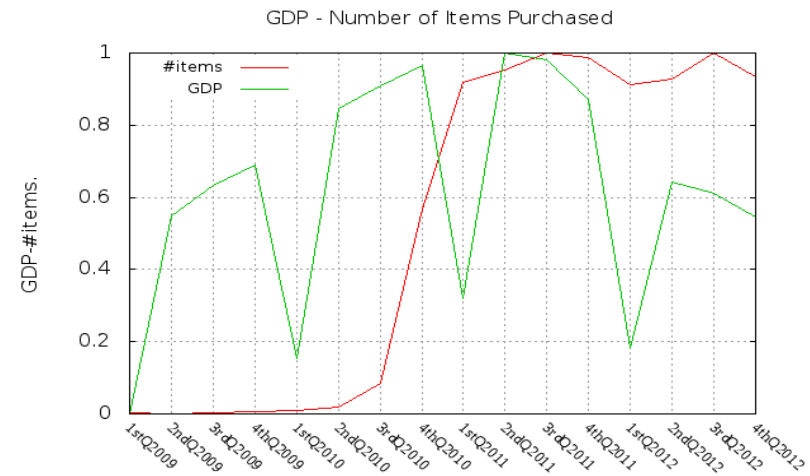
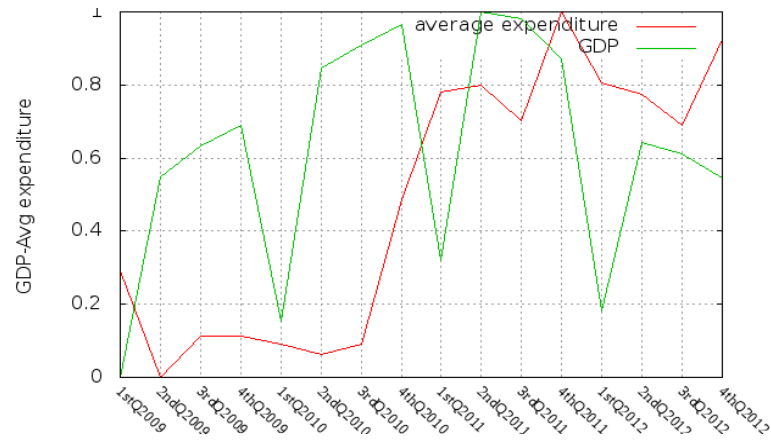
Applicazione spaziale [1]

- each dot is a purchase representative
- if a customer bought products of the same sophistication in different shops, than the distance is weighted with the sophistication



log-normal regression $f(x) = a \log x + b$ $R^2 = 85.72\%$

Misurare il benessere



In prospect ...

- A novel wave of data driven micro-economics indicators based on retail transaction data
- Measuring and nowcasting indicators using retail transaction data
- A natural convergence between data mining/machine learning and complex system modeling

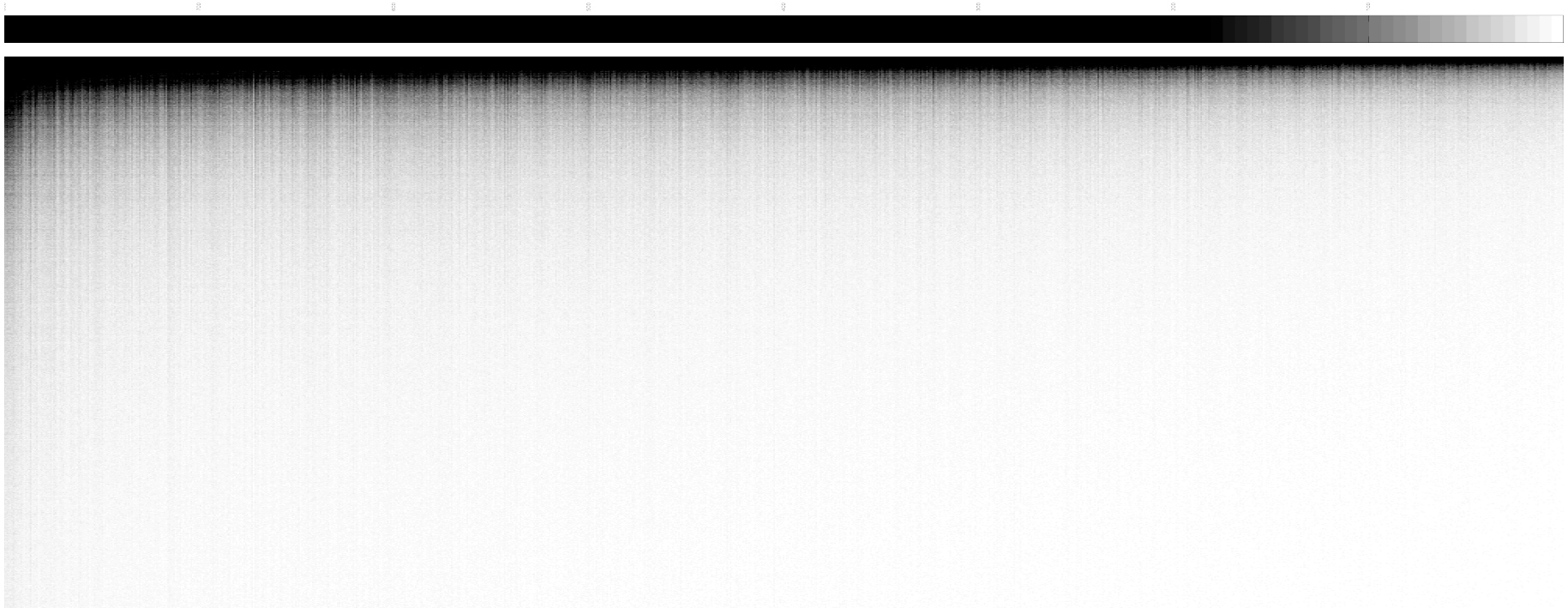
...and in other domains?

UK LastFM Dataset

Users: 76'866

Artists: 247'311

Listenings: 278'740'507



Key references

1. Pennacchioli, D., Coscia, M., Rinzivillo, S., Giannotti, F. and Pedreschi, D., The retail market as a complex system. In EPJ Data Science 3(33), Springer 2014.
2. Pennacchioli, D., Coscia, M., Rinzivillo, S., Pedreschi, D. and Giannotti, F., Explaining the Product Range Effect in Purchase Data. In Proc. BigData Conference, p. 648-656. IEEE 2013.
3. Pennacchioli, D., Coscia, M., and Pedreschi, D., Overlap Versus Partition: Marketing Classification and Customer Profiling in Complex Networks of Products. In ICDE Workshop, 2014.
4. Mirco Nanni and Laura Spinsanti. Forecast analysis for sales in large-scale retail trade. Chapter in *Data Mining in Public and Private Sectors*. IGI Global 2010.
5. Rossetti, G., Pennacchioli, D., Milli, L., Giannotti, F., Pedreschi, D. Predicting Success via Innovators' adoptions. Submitted. 2015

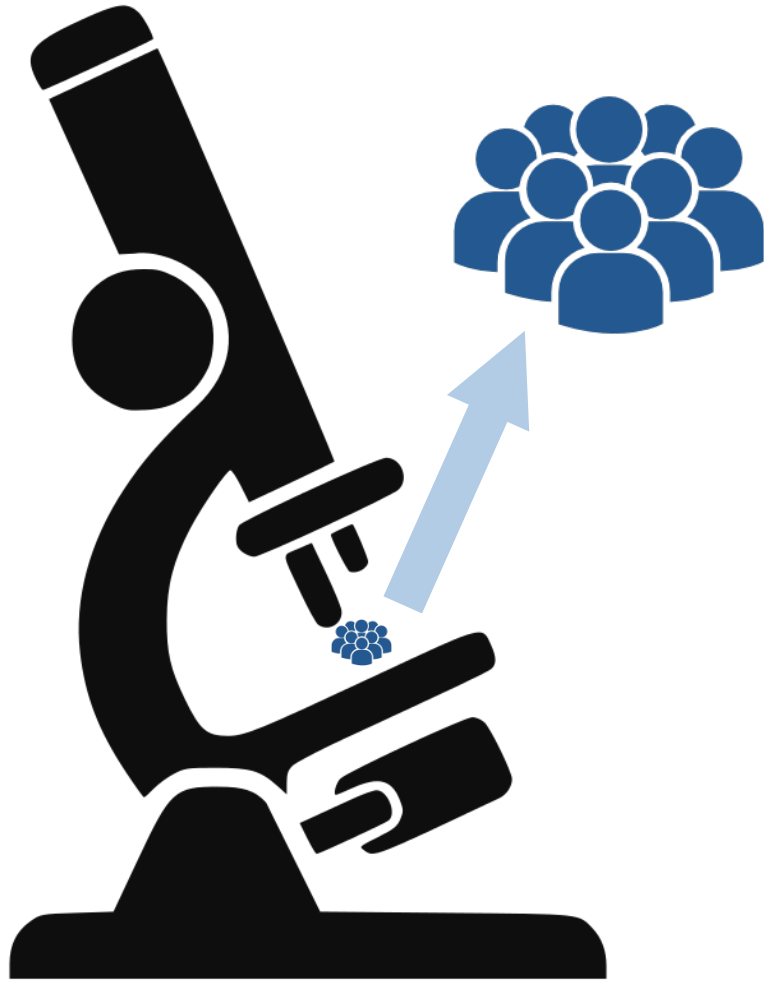
Books

- David Easley, Jon Kleinberg: *Networks, Crowds, and Markets*.
<http://www.cs.cornell.edu/home/kleinber/networks-book/>
- Albert-Laszlo Barabasi. Network Science Book Project
<http://barabasilab.neu.edu/networksciencebook/>
-

SNA course @ Università di Pisa

- <http://didawiki.di.unipi.it/doku.php/wma/start>
- Slides from this course are freely adapted from those of Laszlo Barabasi, Jure Leskovec, Fosca Giannotti, besides my own. Thanks!

Big Data Analytics & Social Mining



a tool to
measure,
understand,
and possibly predict
human behavior

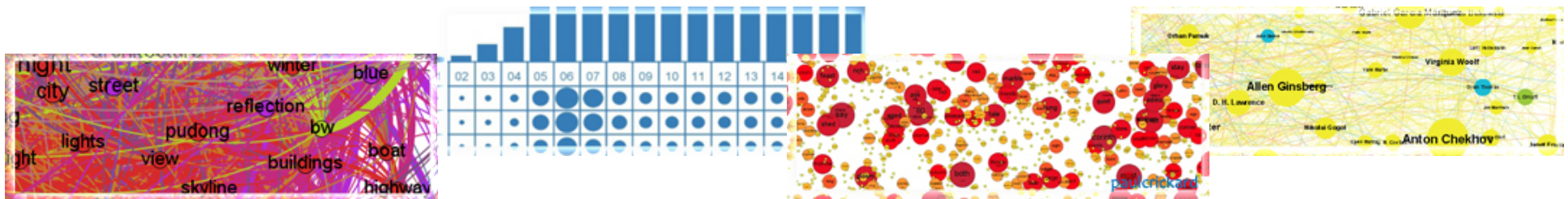


SoBigData Research Infrastructure



Social Mining &
Big Data Analytics

H2020 - www.sobigdata.eu

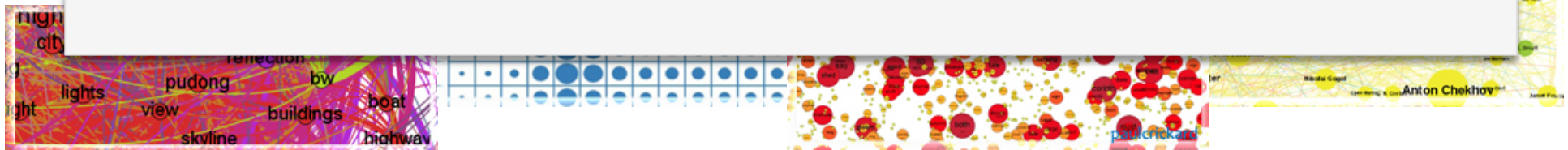




Social mining a OPEN SCIENCE



- An ecosystem of data, methods and competences for serving the community of researchers and innovators ready to exploit the opportunities of big data and to incorporate it in data-driven science and innovation
- open up new research and innovation avenues in multiple fields, including mathematics, ICT, and human, social and economic sciences, by enabling easy comparison, **re-use** and integration of state-of-the-art research data, methods, and services, into new research.





The Consortium



The
University
Of
Sheffield.



UNIVERSITÀ DI PISA



Fraunhofer

FHR



TARTU ÜLIKOOL



INSTITUTE
FOR ADVANCED
STUDIES
LUCCA



Leibniz
Universität
Hannover

KING'S
College
LONDON



SCUOLA
NORMALE
SUPERIORE



Aalto University

ETH Zürich



TU Delft
Delft
University of
Technology

Existing national RI's to be integrated



general architecture
for text engineering



Fraunhofer
IGD



L3S Research Center





Activities



Economics & Finance



Scalable Data Analytics



Human Mobility Analytics



Opinion Mining



Privacy, Security & Trust



Search & Web Analytics



Social Network Analysis



Social Sensing

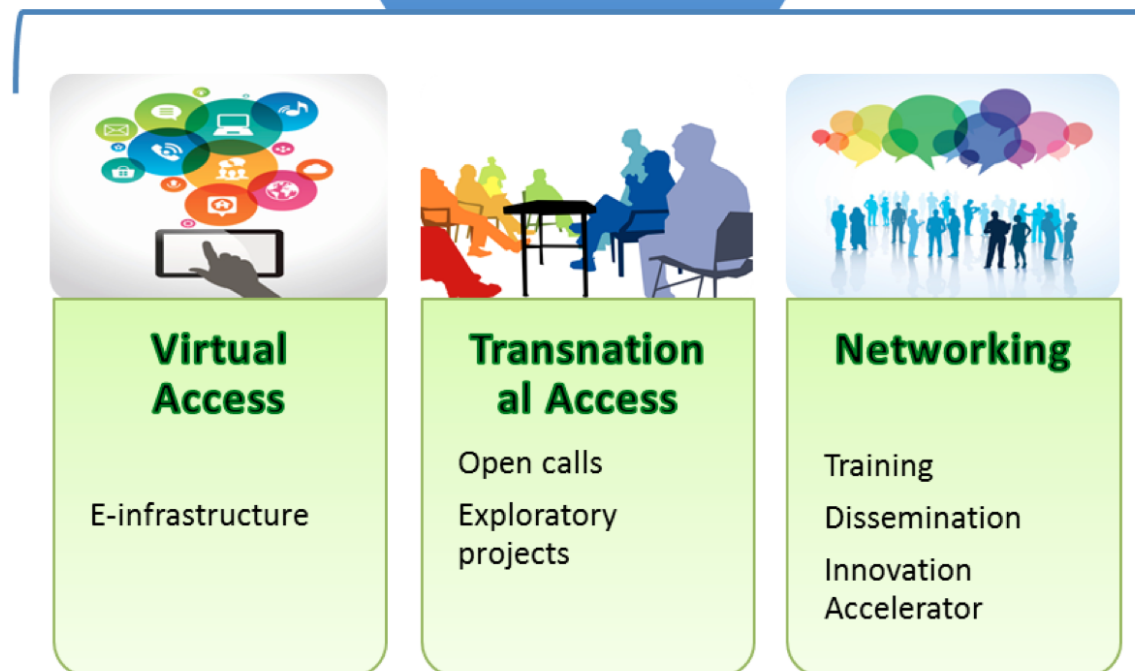
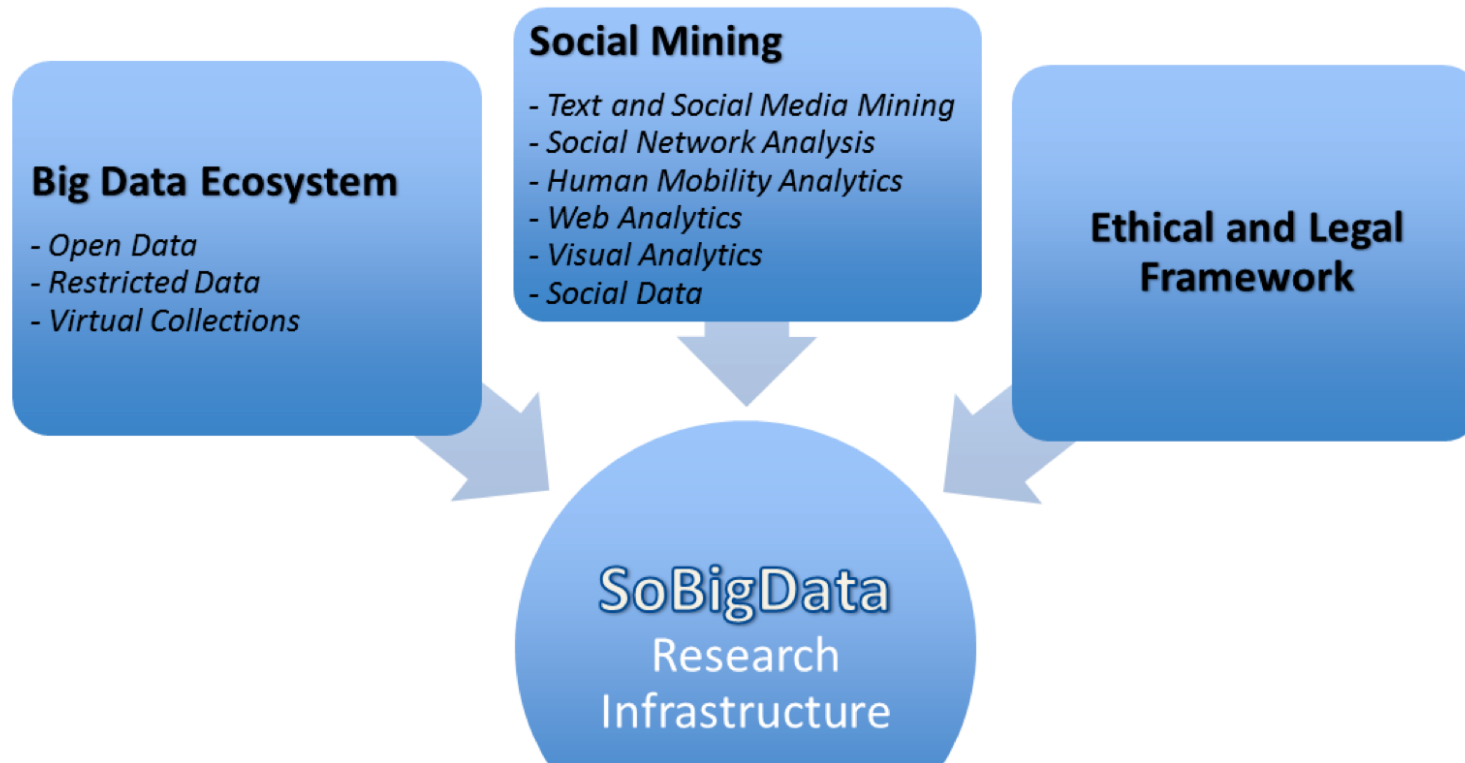


Well-Being



Thematic Clusters





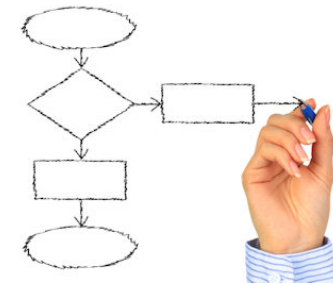


SoBigData Access

- Transnational Access
 - Exploratory Projects
 - Blue-sky projects



- Virtual Access
 - Data and Methods Catalog
 - Modular virtual research environment



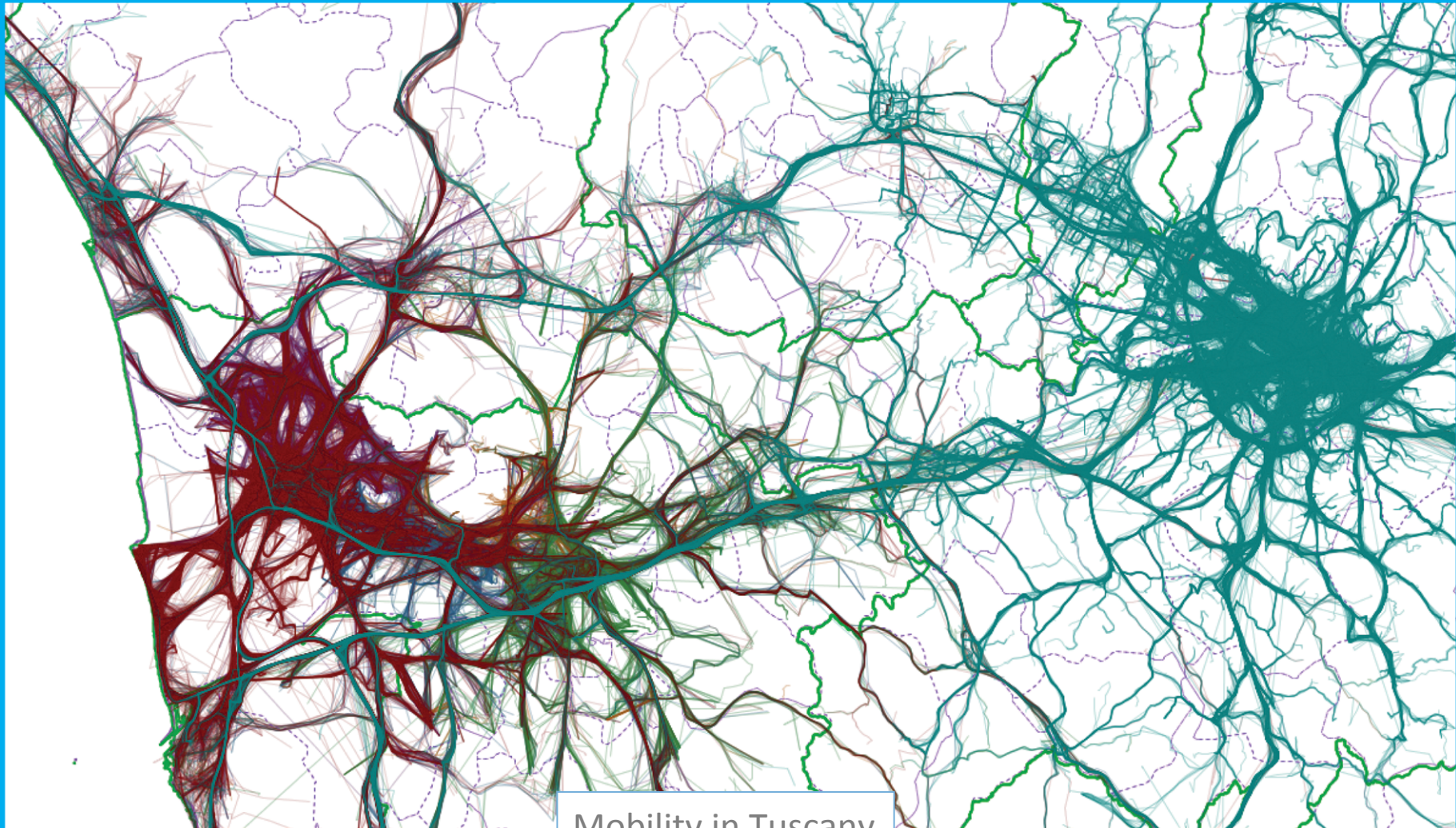


Exploratories

Social Mining Research Environments
tailored on specific multidisciplinary
domains



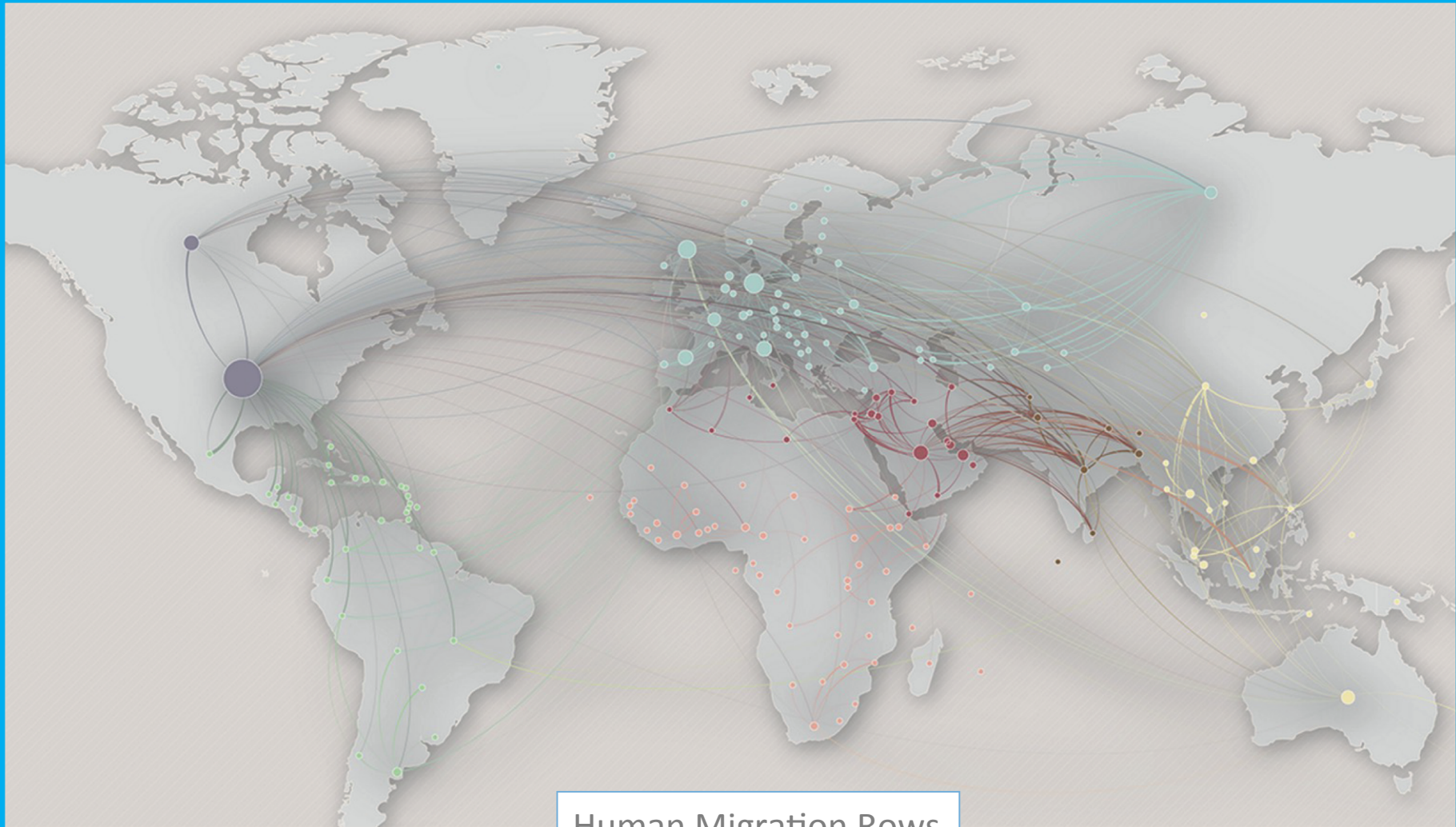
Exploratory: **Big Data** for **Human Mobility**



Mobility in Tuscany



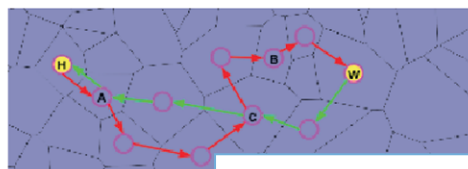
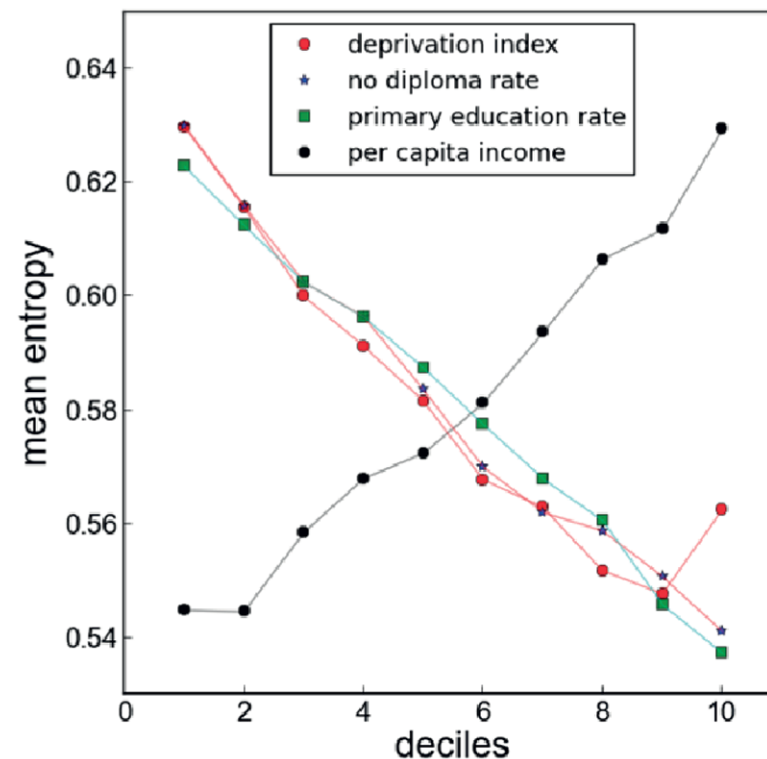
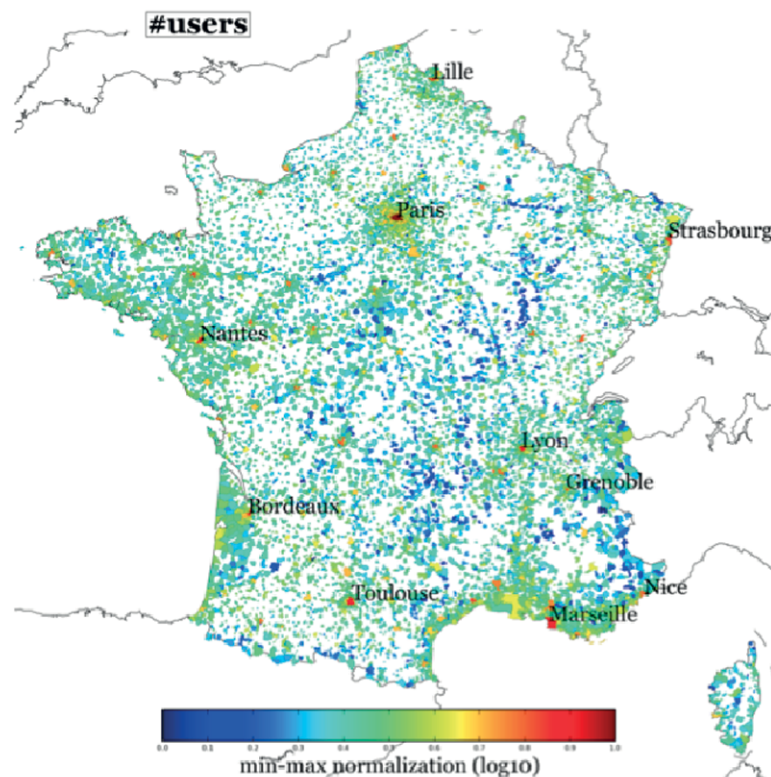
Exploratory: **Big Data** for **Migration Studies**



Human Migration Rows



Exploratory: Big Data for Well Being



$$d_i^{(n)} = \sum_{j=1}^{|V|} \frac{1}{k_j} M_{ij} p_j^{(n-1)} \forall i$$

$$p_j^{(n)} = \sum_{i=1}^{|U|} \frac{1}{k_i} M_{ij} d_i^{(n-1)} \forall j$$

Deprivation Index (in France) predicted with Mobile Phone traces



Exploratory: **Big Data for Developing Countries**



Origin Destination Flows estimated with Mobile Phone in Ivory Cote



Exploratory: Big Data for Societal Debates



Polarization, controversy and topic trends on societal debates through social media



The SoBigData Stakeholders

- **Big data analysts and social informatics researchers**, who want to enhance their algorithms to deal with social data, gain multi-disciplinary research skills, harmonise existing data and analytics infrastructures, and engage other research communities in the development of these key enabling technologies for the future digital economy and society;
- **Economists, social science and humanities researchers, journalists, policy and law makers**, who have to analyse the avalanche of (big) social data, in order to gain insight and actionable knowledge.
- **Researchers in related communities**, who would like to use the algorithms, the analytical competences and data infrastructure;
- **Industrial innovators & startupper**s, who would like to create rapid proof-of-concepts of data-driven innovative ideas and services;
- **The public as a whole**, who would like to understand their role in the production, consumption and value-creating of social data.



SoBigData
it



GENNAIO
29
2016



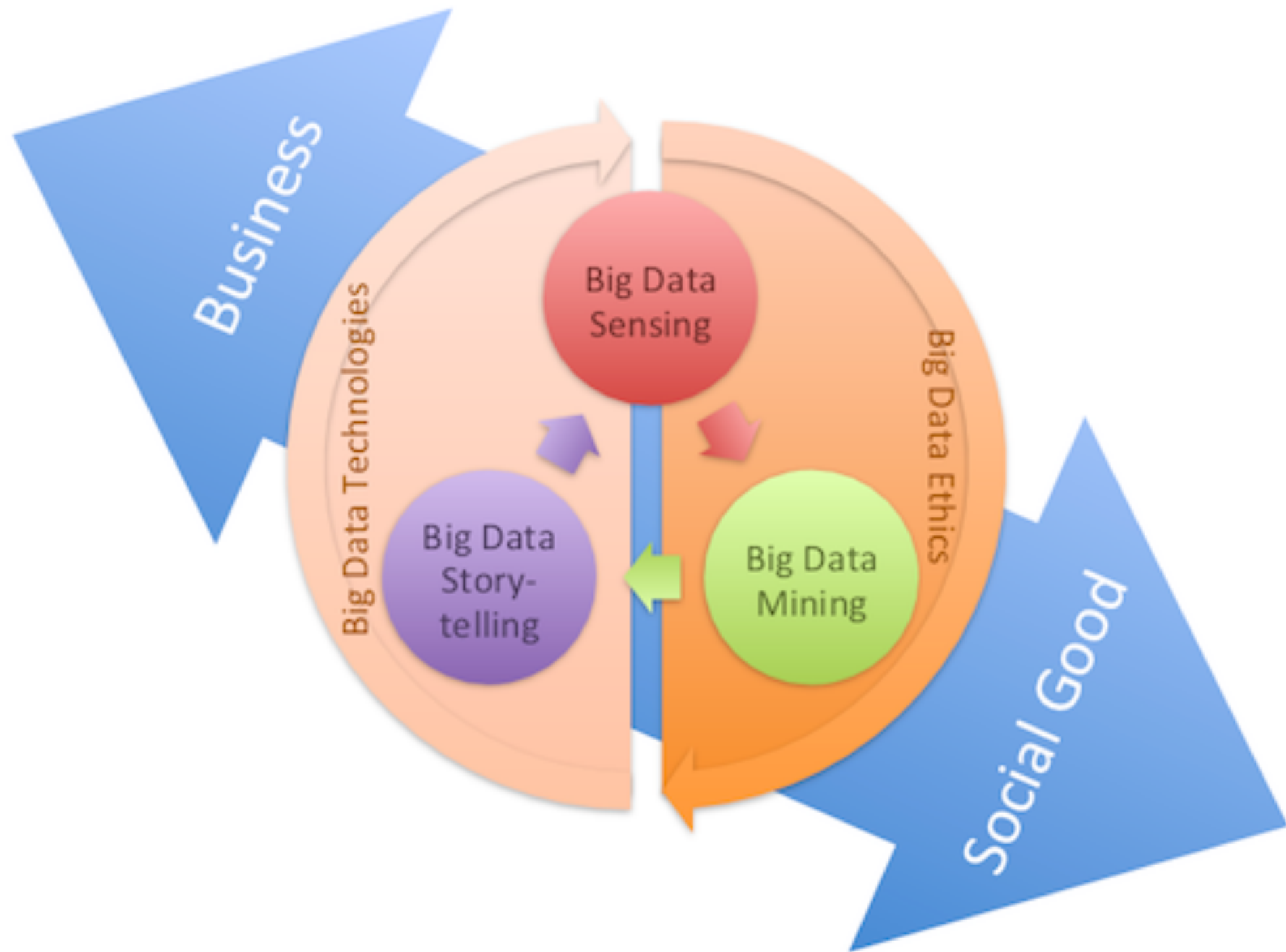
MASTER BIG DATA 2016



Data scientist



- ... a new kind of professional has emerged, the **data scientist**, who combines the skills of **software programmer**, **statistician** and **storyteller/artist** to extract the nuggets of gold hidden under mountains of data.





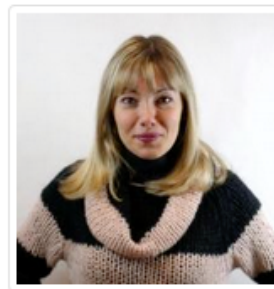
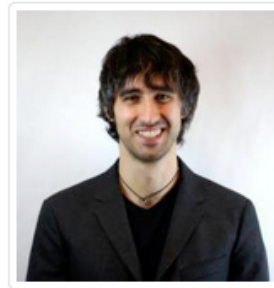
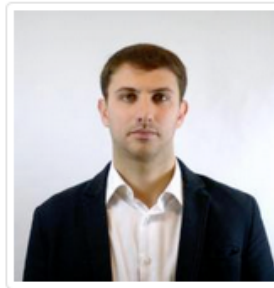
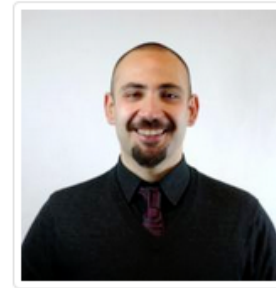
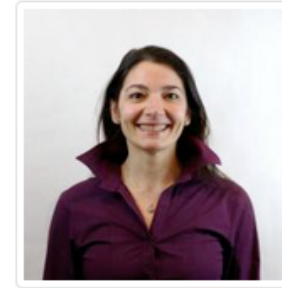
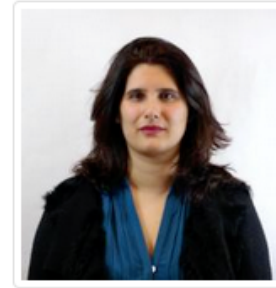
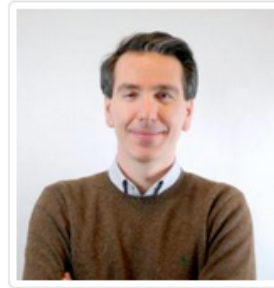
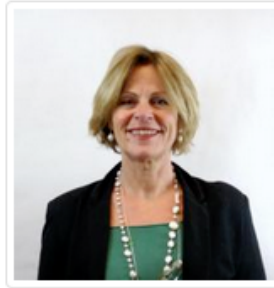
www.sobigdata.eu



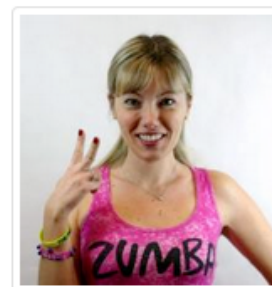
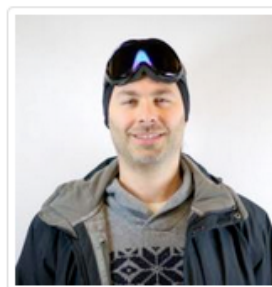
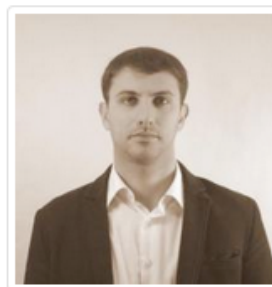
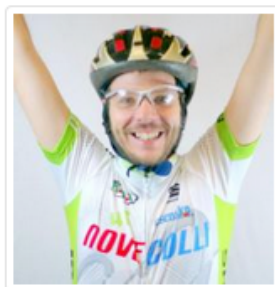
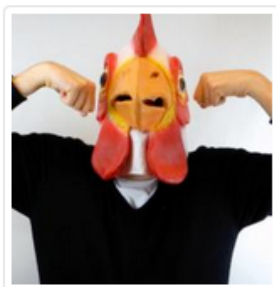
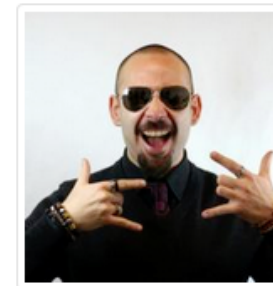
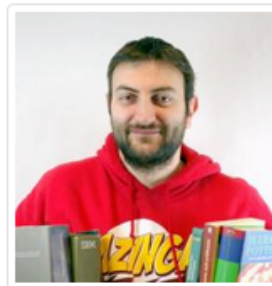
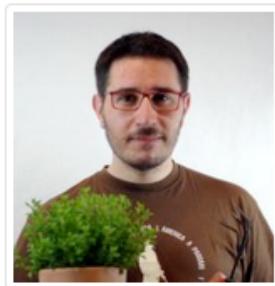
fosca.giannotti@isti.cnr.it

This is the work of many people for a long time

- Salvo Rinzivillo, Mirco Nanni, Roberto Trasarti, Salvatore Ruggieri, Chiara Renso, Anna Monreale, Franco Turini
- all the fantastic folks at KDD LAB Pisa
- many international collaborators
- Thanks!



**Knowledge Discovery
& Data Mining Lab**
<http://kdd.isti.cnr.it>



**Knowledge Discovery
& Data Mining Lab**
<http://kdd.isti.cnr.it>

Key publications

- F Giannotti, M Nanni, F Pinelli, D Pedreschi. Trajectory pattern mining. ACM SIGKDD 2007
- F Giannotti, D Pedreschi. Mobility, data mining and privacy: Geographic knowledge discovery. Springer, 2008
- A Monreale, F Pinelli, R Trasarti, F Giannotti. WhereNext: a location predictor on trajectory pattern mining. ACM SIGKDD 2009
- S Rinzivillo, D Pedreschi, M Nanni, F Giannotti, N Andrienko, G Andrienko. Visually driven analysis of movement data by progressive clustering. Information Visualization 7 (3-4), 225-239. 2008
- D Wang, D Pedreschi, C Song, F Giannotti, AL Barabasi. Human mobility, social ties, and link prediction. ACM SIGKDD 2011
- F Giannotti, M Nanni, D Pedreschi, F Pinelli, C Renso, S Rinzivillo, R Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. The VLDB Journal 20(5) 2011
- R Trasarti, F Pinelli, M Nanni, F Giannotti. Mining mobility user profiles for car pooling. ACM SIGKDD 2011
- M Coscia, G Rossetti, F Giannotti, D Pedreschi. Demon: a local-first discovery method for overlapping communities. ACM SIGKDD 2012
- D Pennacchioli, M Coscia, S Rinzivillo, F Giannotti, D Pedreschi. The retail market as a complex system. EPJ Data Science 3 (1), 1-27 (2014)
- A Monreale, S Rinzivillo, F Pratesi, F Giannotti, D Pedreschi. Privacy-by-design in big data analytics and social mining. EPJ Data Science 3 (1), 1-26 (2014)
- Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti & Albert-László Barabási. Returners and explorers dichotomy in human mobility. Nature Communications 6, Article number: 8166 (2015) doi:10.1038/ncomms9166 (2015)

Key publications

- M Coscia, G Rossetti, F Giannotti, D Pedreschi. Demon a local-first discovery method for overlapping communities. ACM SIGKDD 2012
- S Rinzivillo, S Mainardi, F Pezzoni, M Coscia, D Pedreschi, F Giannotti. Discovering the geographical borders of human mobility. KI-Künstliche Intelligenz 26 (3) 2012
- D Pennacchioli, M Coscia, S Rinzivillo, D Pedreschi, F Giannotti. Explaining the Product Range Effect in Purchase Data. IEEE BIGDATA 2013
- B Furletti, L Gabrielli, C Renso, S Rinzivillo. Analysis of GSM Calls Data for Understanding User Mobility Behavior. IEEE BIGDATA 2013
- L Milli, A Monreale, G Rossetti, D Pedreschi, F Giannotti, F Sebastiani. Quantification trees. IEEE ICDM 2013
- Giusti, Marchetti, Pratesi, Salvati, Pedreschi, Giannotti, Rinzivillo, Pappalardo, Gabrielli. Small area model based estimators using Big Data Sources. Journal of Official Statistics, 31(2) 2015.
- Furletti, Gabrielli, Garofalo, Giannotti, Milli, Nanni, Pedreschi, Vivio. Use of mobile phone data to estimate mobility flows. Measuring urban population and intercity mobility using big data in an integrated approach. Italian Symposium on Statistics, 2014.
- Luca Pappalardo, Maarten Vanhoof, Zbigniew Smoreda, Dino Pedreschi, Fosca Giannotti. Human Mobility and Economic Development. IEEE BIG DATA (2015).

Vision papers

- F Giannotti, D Pedreschi, A Pentland, P Lukowicz, D Kossmann, J Crowley, D Helbing. **A planetary nervous system for social mining and collective awareness.** The European Physical Journal Special Topics 214 (1), 49-75, 2012
- J van den Hoven, D Helbing, D Pedreschi, J Domingo-Ferrer, F Giannotti . **FuturICT—The road towards ethical ICT.** The European Physical Journal Special Topics 214 (1), 153-181, 2012
- M Batty, KW Axhausen, F Giannotti, A Pozdnoukhov, A Bazzani, M Wachowicz. **Smart cities of the future.** The European Physical Journal Special Topics 214 (1), 481-518, 2012