Saïd Business School
UNIVERSITY OF OXFORD

WWW.SBS.OXFORD.EDU
EDUCATING LEADERS FOR 800 YEARS

Cambridge Networks Day 2013
Sainsbury Laboratory
University of Cambridge

# Social Mechanisms in Real and Virtual Networks

Felix Reed-Tsochas
CABDyN Complexity Centre & Oxford Martin School
7 May 2013

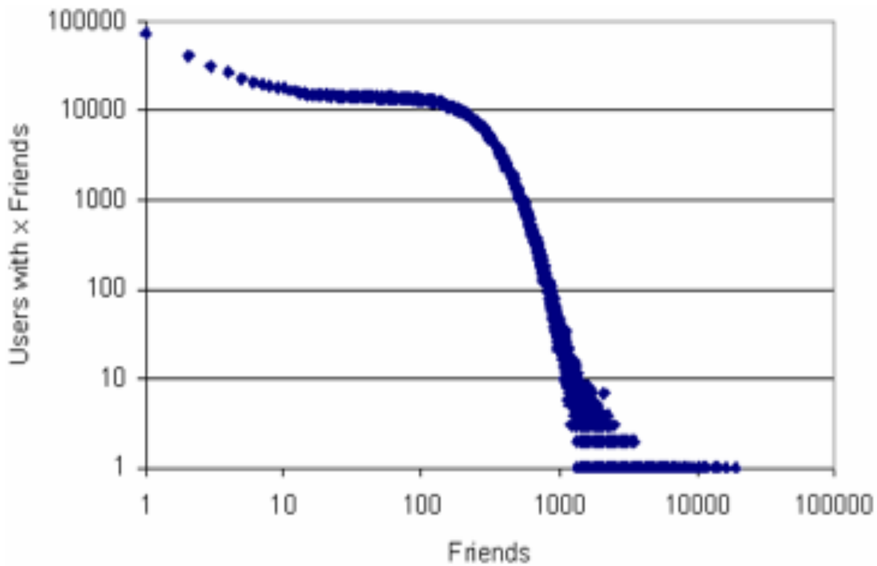
Cabdyn Complexity Centre

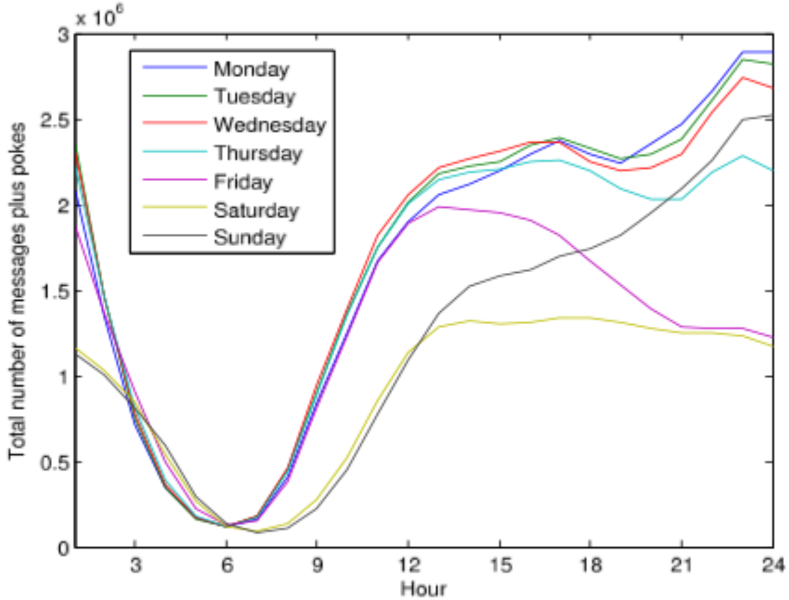
# Outline

- Motivation for studying the spread of Facebook applications
  - Online social networks
  - Markets for cultural goods
  - Diffusion of innovations

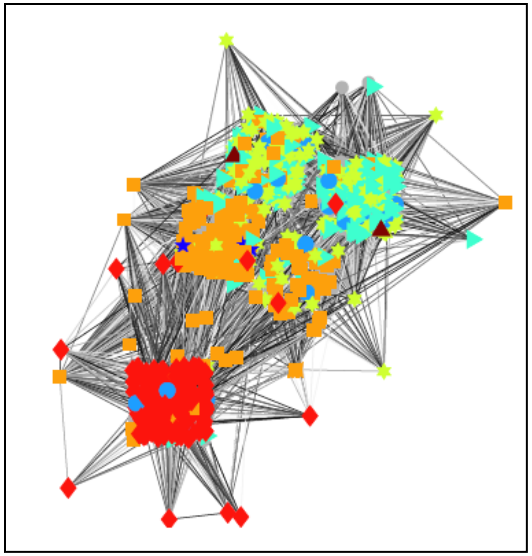
    (spatial, network)
- Online environment: local and global information
- Empirical analysis and temporal fluctuation scaling
- Online experiments and microscopic models
- The social brain hypothesis and ego-network structure

# Online social networks




Golder, Wilkinson and Huberman (2006)
Rhythms of social interaction:
messaging within a massive online network
*arXiv:cs/0611137v1*



Traud, Mucha and Porter (2012)
Social structure of Facebook networks
*Physica A*, **391**, 4165–4180.

# Social influence and cultural products


Hit songs, books, and movies are many times more successful than average, suggesting that "the best" alternatives are qualitatively different from "the rest"; yet experts routinely fail to predict which products will succeed. We investigated this paradox experimentally, by creating an artificial "music market" in which 14,341 participants downloaded previously unknown songs either with or without knowledge of previous participants' choices. Increasing the strength of social influence increased both inequality and unpredictability of success. Success was also only partly determined by quality: The best songs rarely did poorly, and the worst rarely did well, but any other result was possible.

Salganik, Dodds and Watts (2006) Experimental study of inequality and unpredictability in an artificial cultural market, *Science* **311**, 854-856.

# Inequality of success

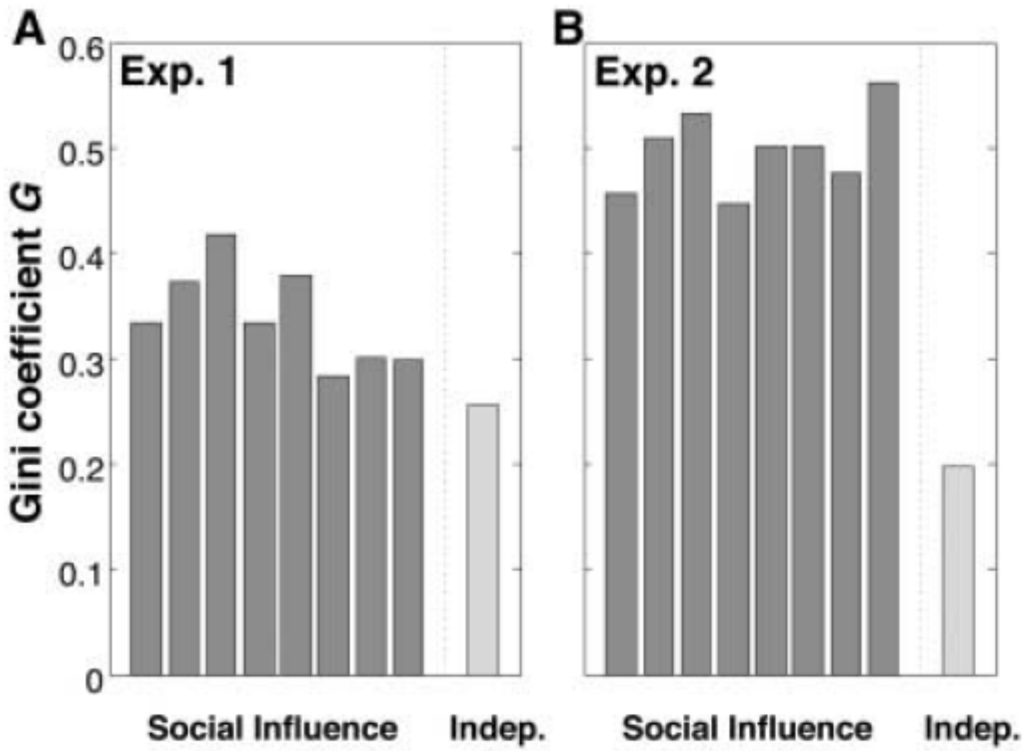



**Fig. 1.** Inequality of success for social influence (dark bars) and independent (light bars) worlds for (**A**) experiment 1 and (**B**) experiment 2. The success of a song is defined by $m_i$, its market share of downloads ($m_i = d_i / \sum_{k=1}^{S} d_k$, where $d_i$ is song $i$'s download count and $S$ is the number of songs). Success inequality is defined by the Gini coefficient $G = \sum_{i=1}^{S} \sum_{j=1}^{S} |m_i - m_j| / 2S \sum_{k=1}^{S} m_k$, which represents the average difference in market share for two songs normalized to fall between 0 (complete equality) and 1 (maximum inequality). Differences between independent and social influence conditions are significant ($P < 0.01$) (*18*).


Salganik, Dodds and Watts (2006) Experimental study of inequality and unpredictability in an artificial cultural market, *Science* **311**, 854-856.

# Unpredictability of success

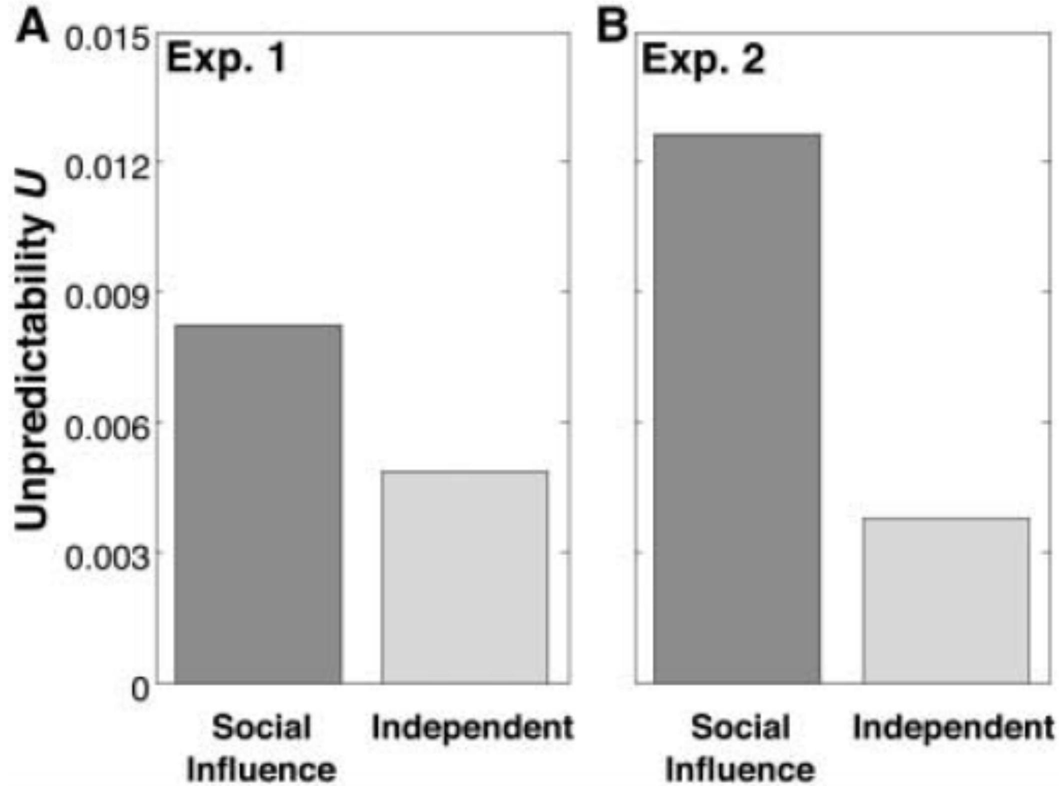



**Fig. 2.** Unpredictability of success for (**A**) experiment 1 and (**B**) experiment 2. In both experiments, success in the social influence condition was more unpredictable than in the independent condition. Moreover, the stronger social signal in experiment 2 leads to increased unpredictability. The measure of unpredictability $u_i$ for a single song $i$ is defined as the average difference in market share for that song between all pairs of realizations; i.e., $u_i = \sum_{j=1}^{W} \sum_{k=j+1}^{W} |m_{i,j} - m_{i,k}| / \binom{W}{2}$, where $m_{i,j}$ is song $i$'s market share in world $j$ and $W$ is the number of worlds. The overall unpredictability measure $U = \sum_{i=1}^{S} u_i / S$ is then the average of this measure over all $S$ songs. For the independent condition, we randomly split the single world into two subpopulations to obtain differences in market shares, and we then averaged the results over 1000 of these splits. All differences are significant ($P < 0.01$) (*18*).

Salganik, Dodds and Watts (2006) Experimental study of inequality and unpredictability in an artificial cultural market, *Science* **311**, 854-856.

# Cultural markets

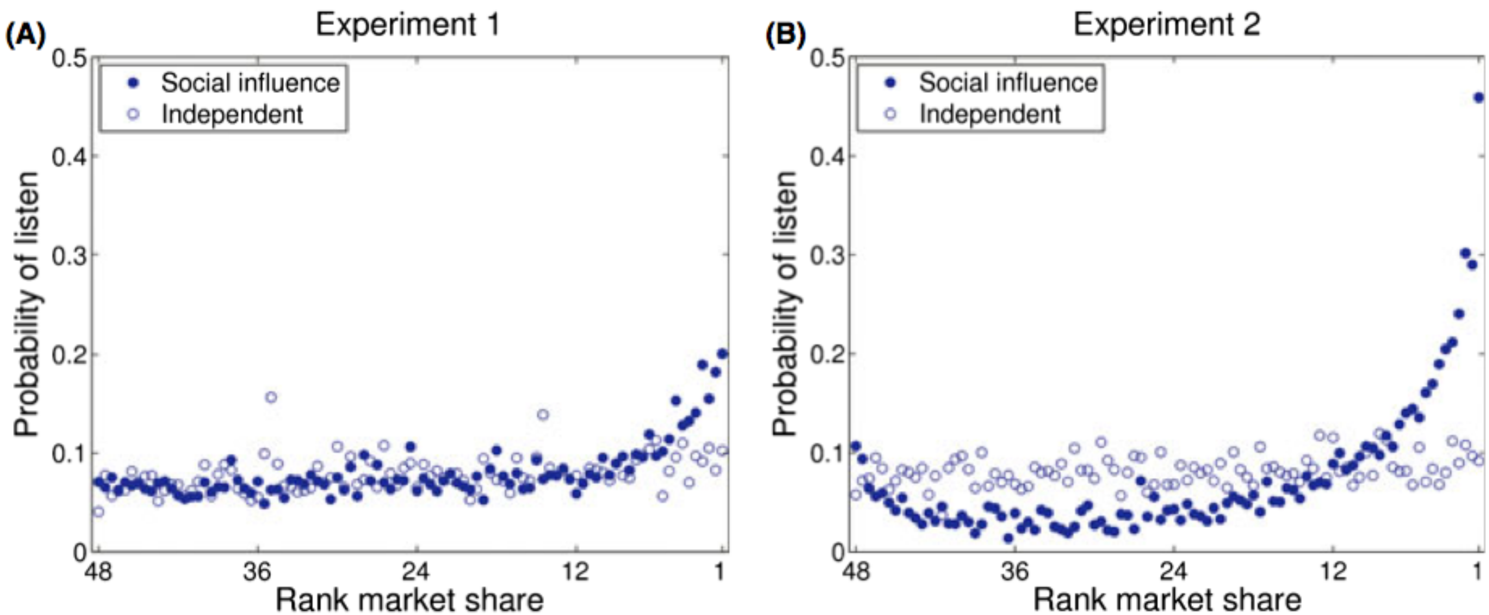
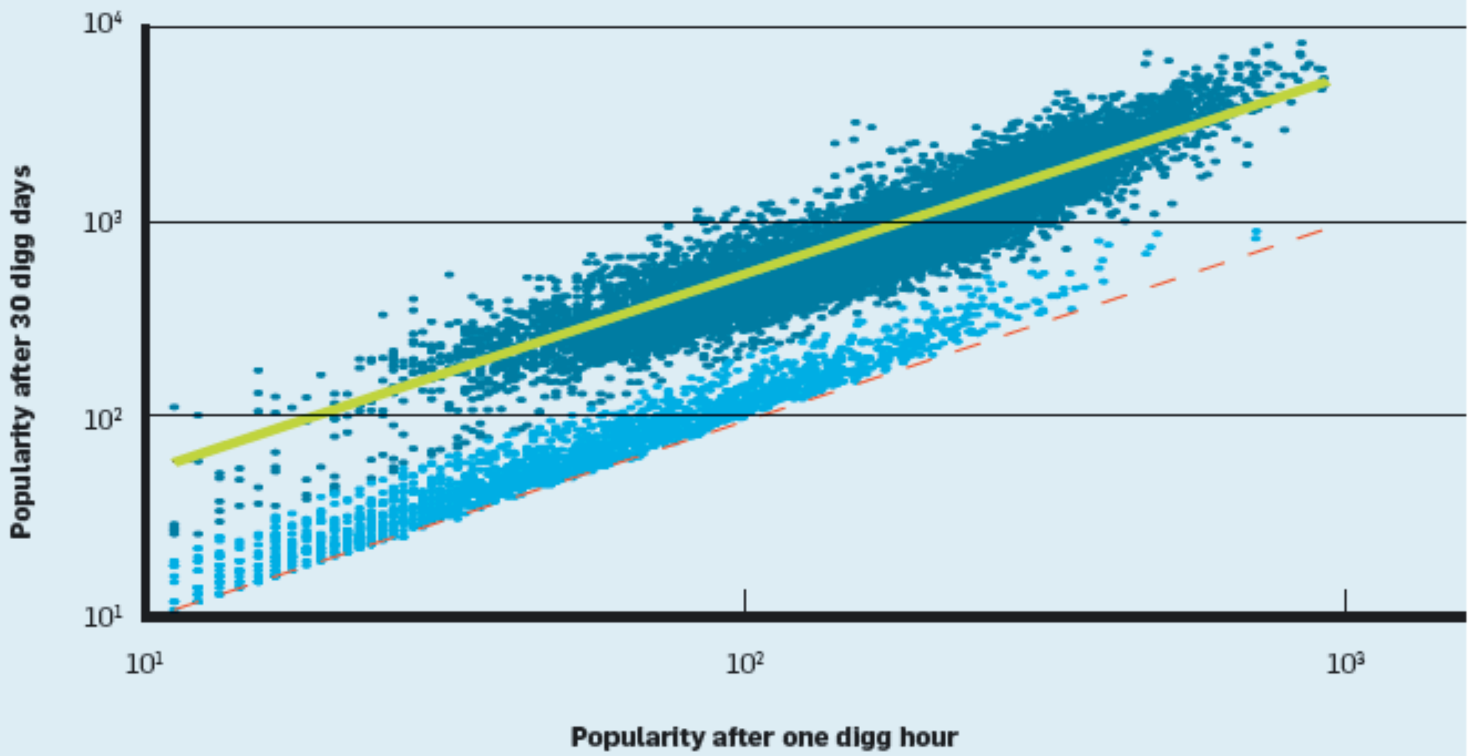



Fig. 3. Probability of listening to a song of a given market rank in Experiment 1(A) and Experiment 2(B). Participants in Experiment 2 were more likely to listen to more popular songs.

Salganik and Watts (2009)
Web-based experiments for the study of collective social dynamics in cultural markets, *Cognitive Science* **1**, 439-468.

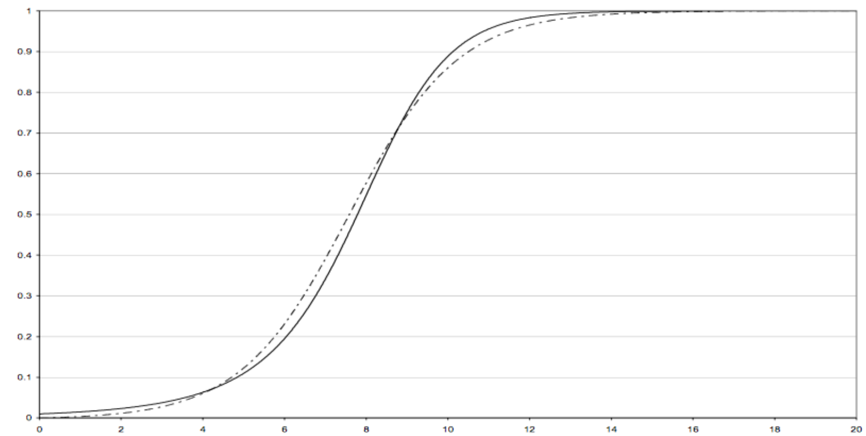
# Predicting online popularity

**Figure 3. Correlation of digg counts on the 17,097 promoted stories in the data set older than 30 days. A k-means clustering separates 89% of the stories into an upper cluster; the other stories are a lighter shade of blue. The bold line indicates a linear fit with slope 1 on the upper cluster, with a prefactor of 5.92 (Pearson correlation coefficient of 0.90).**




Szabo and Huberman (August 2010). Predicting the popularity of online content, *Communications of the ACM* **53**:80-88.

# Innovation diffusion




Classic examples:

Switch to hybrid corn by US farmers (Griliches 1957)

Antibiotic prescriptions spreading by word-of-mouth between physicians

(Coleman, Katz & Menzel 1957)

Some methodological challenges:

Incomplete sampling and sampling biases

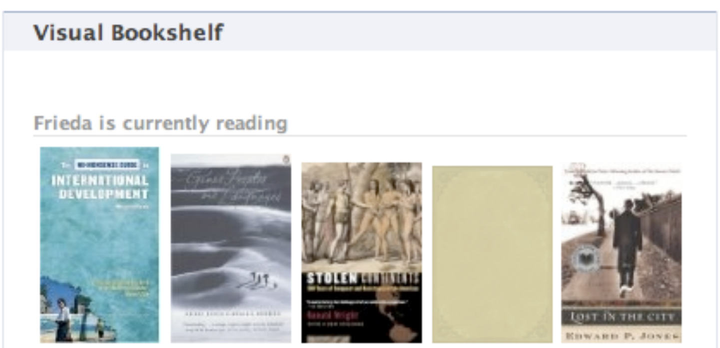
Recent re-analyses suggest that effect of sales reps etc. has been neglected

Difficult to control for external drives (e.g. advertising, media)

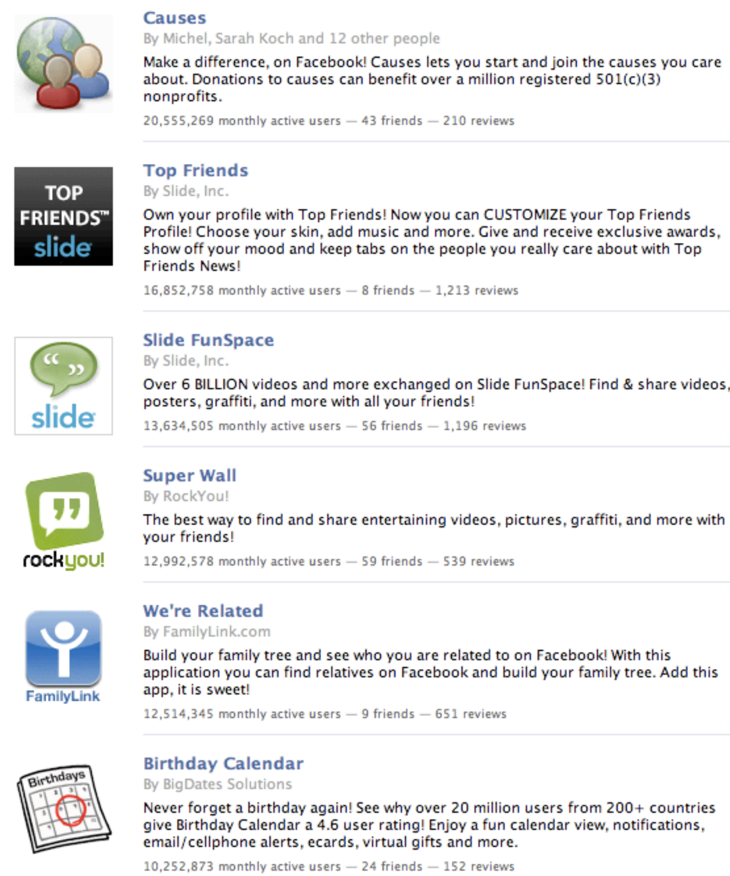
Young (2009), Innovation diffusion in heterogeneous populations: contagion, social influence, and social learning, *American Economic Review* **99**: 1899-1924.

# The Facebook environment

## Local information


## Global information




[Note: These are from a more recent version of Facebook]

# Information and influence on Facebook

Local info – Facebook informed FB friends of application installations, and users could look up which applications their FB friends had installed.

Global info – Users could access a rank ordered list of all applications, giving the overall number of installations for each, i.e. a real-time “best seller” list.

Potential constraints – Applications are free, but too many clog up a user’s FB page.

Local popularity – Friends may have similar interests and tastes (i.e. homophily)

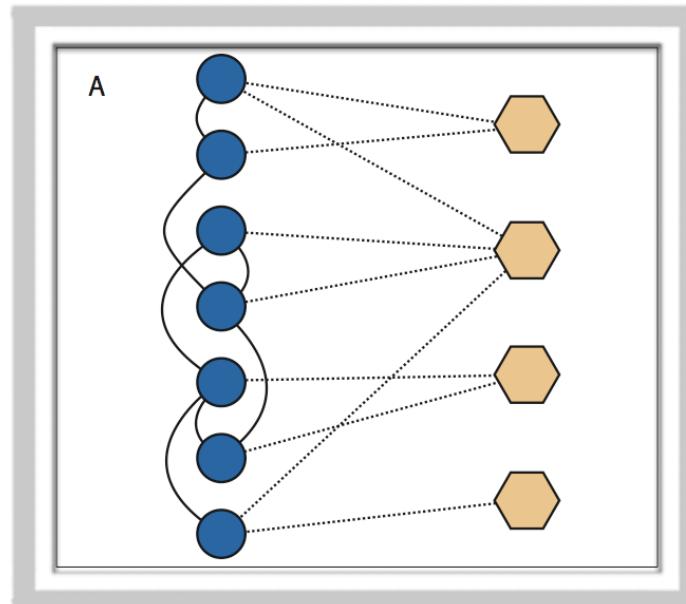
Global popularity – A high ranking may:

(i) lower search costs

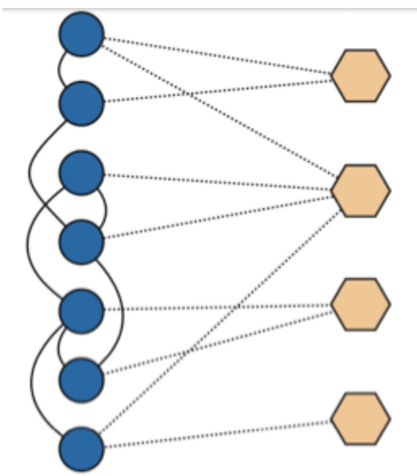
(ii) signal high quality

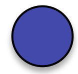
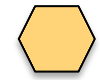
(iii) signal superior functionality

Facebook users and applications

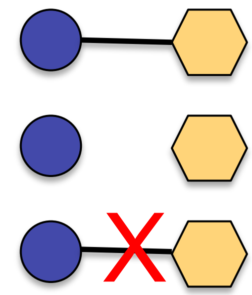


# Measuring social influence



User $j$  Application $i$ 

$$S_{i,j}(t) = \begin{cases} 1 & \text{user } j \text{ adopts application } i \\ 0 & \text{user } j \text{ does nothing} \\ -1 & \text{user } j \text{ drops application } i \end{cases}$$



Net activity $$f_i(t) = n_i(t) - n_i(t-1) = \sum_{j=1}^{N} S_{i,j}(t) = \sum_{k=1}^{N-n_i(t)} S_{i,j_k}(t)$$

Mean of time activity series $$\mu_i \equiv \langle f_i \rangle = \frac{1}{T_i} \sum_{t=1}^{T_i} f_i(t)$$

SD of time activity series $$\sigma_i = \left( \frac{1}{T_i - 1} \sum_{t=1}^{T_i} \left[ f_i(t) - \langle f_i \rangle \right]^2 \right)^{1/2}$$

# Fluctuation scaling

Scaling properties of fluctuations in complex systems [Taylor's law]

*fluctuations* $\approx$ *const. x average*$^{\alpha}$

Decompose additive quantity $f_i$ (where i denotes signal or measurement) into random variables $V_{i,n}^{\Delta t}(t)$ for some finite duration $[t, t+\Delta t)$

$$f_i^{\Delta t}(t) = \sum_{n=1}^{N_i^{\Delta t}(t)} V_{i,n}^{\Delta t}(t)$$

e.g. $N_i^{\Delta t}(t)$ – number of transactions with shares in company $i$

$V_{i,n}^{\Delta t}(t)$ – value of the $n^{th}$ transaction

$f_i^{\Delta t}(t)$ – total trading activity of stock $i$

Eisler, Bartos, Kertesz (2008). Fluctuation scaling in complex systems: Taylor's law and beyond, *Advances in Physics* **57**: 89-142.

# Temporal fluctuation scaling

If we assume that $V_{i,n}^{\Delta t}(t) \geq 0$ so that the time average of $f_i^{\Delta t}$ doesn't vanish, then we can write it as:

$$\langle f_i^{\Delta t}(t) \rangle = \frac{1}{Q}\sum_{q=0}^{Q-1} f_1^{\Delta t}(q\Delta t) = \frac{1}{Q}\sum_{q=0}^{Q-1}\sum_{n=1}^{N_i^{\Delta t}(q\Delta t)} V_{i,n}^{\Delta t}(q\Delta t)$$

Where $Q=T/\Delta t$ and $T$ is the total time of measurement.

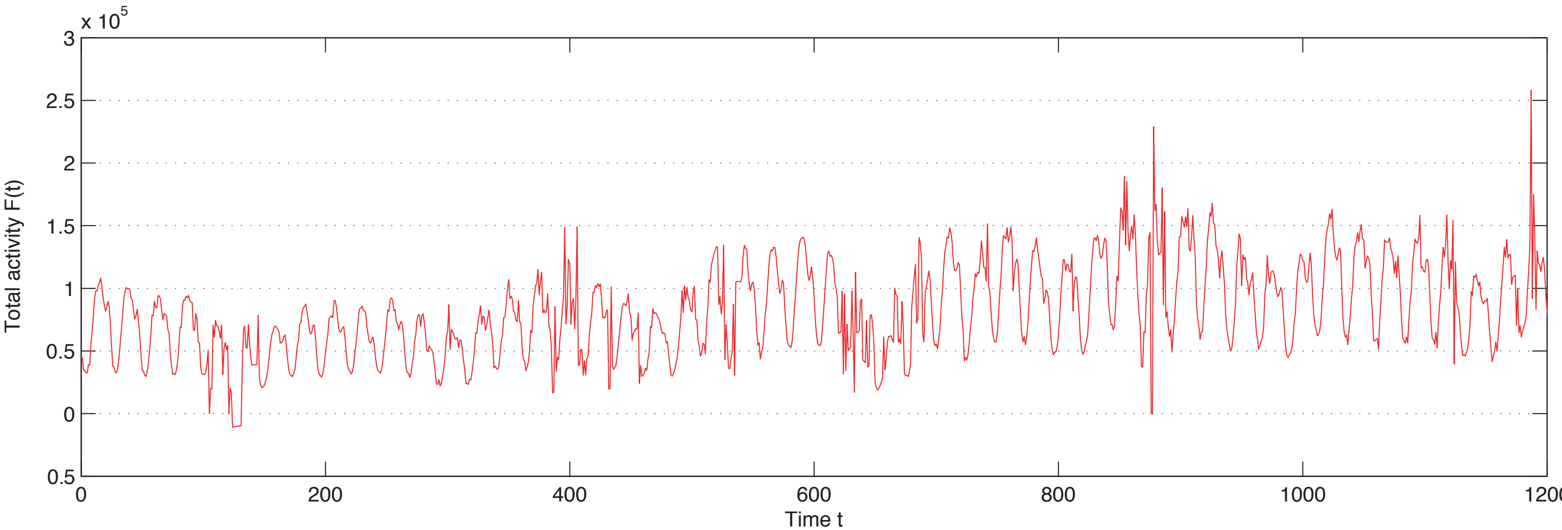
On any time scale the variance can be obtained as a time average:

$$\sigma_i^2(\Delta t) = \langle \left[ f_i^{\Delta t} \right]^2 \rangle - \langle f_i^{\Delta t} \rangle^2$$

If $f$ is positive and additive we frequently observe:

$$\sigma_i(\Delta t) \propto \langle f_i \rangle^{\alpha_T}$$

# Time-series activity data



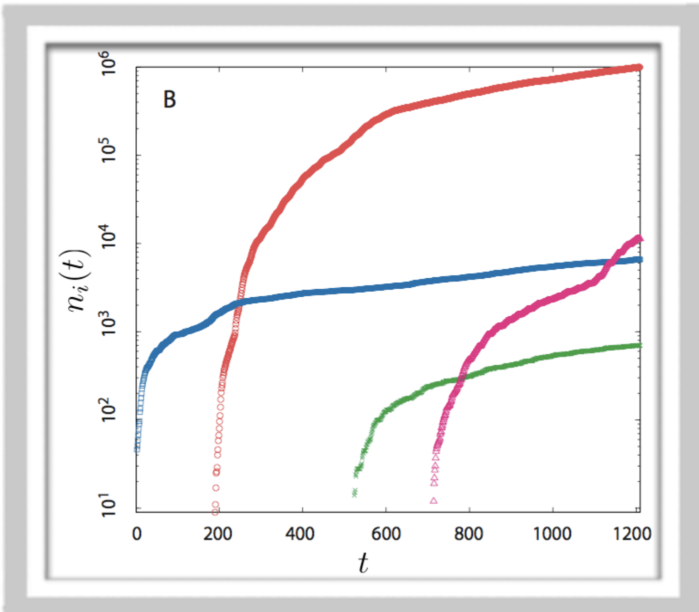

$$f_i(t) = n_i(t) - n_i(t-1) = \sum_{j=1}^{N} S_{i,j}(t)$$

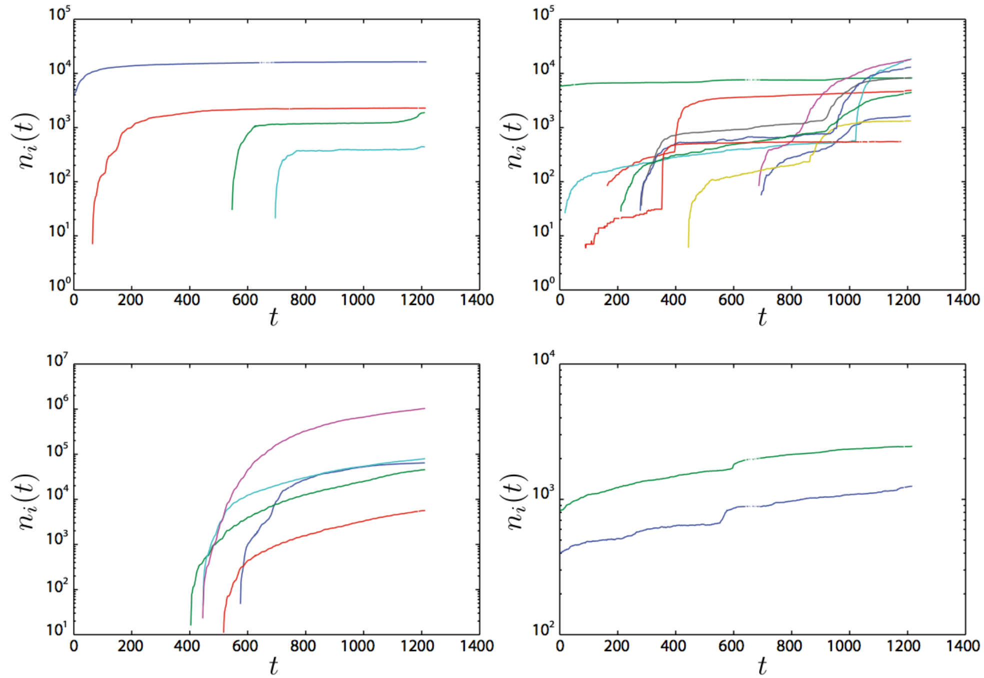
"Movies" application

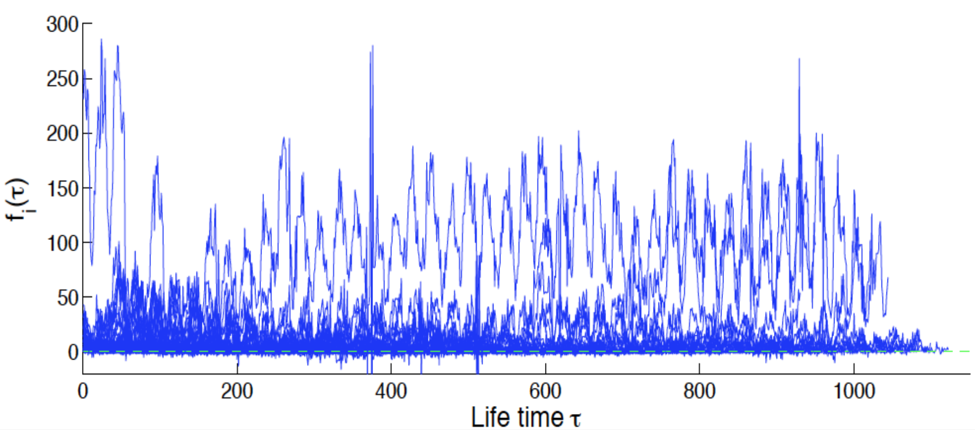
| | |
|---|---|
| 7/4/07 0:01 | 1820 |
| 7/4/07 1:01 | 1836 |
| 7/4/07 2:01 | 1839 |
| 7/4/07 3:01 | 1847 |
| 7/4/07 4:01 | 1852 |
| 7/4/07 5:01 | 1860 |
| 7/4/07 6:01 | 1867 |
| 7/4/07 7:01 | 1874 |
| 7/4/07 8:01 | 1880 |
| 7/4/07 9:03 | 1889 |
| 7/4/07 10:01 | 1899 |
| 7/4/07 11:02 | 1908 |
| 7/4/07 12:02 | 1921 |
| 7/4/07 13:02 | 1931 |
| 7/4/07 14:01 | 1949 |
| 7/4/07 15:01 | 1964 |
| 7/4/07 16:01 | 1987 |
| 7/4/07 17:03 | 2000 |
| 7/4/07 18:03 | 2014 |
| 7/4/07 19:03 | 2025 |
| 7/4/07 20:03 | 2036 |
| 7/4/07 21:03 | 2048 |
| 7/4/07 22:03 | 2060 |
| 7/4/07 23:03 | 2071 |

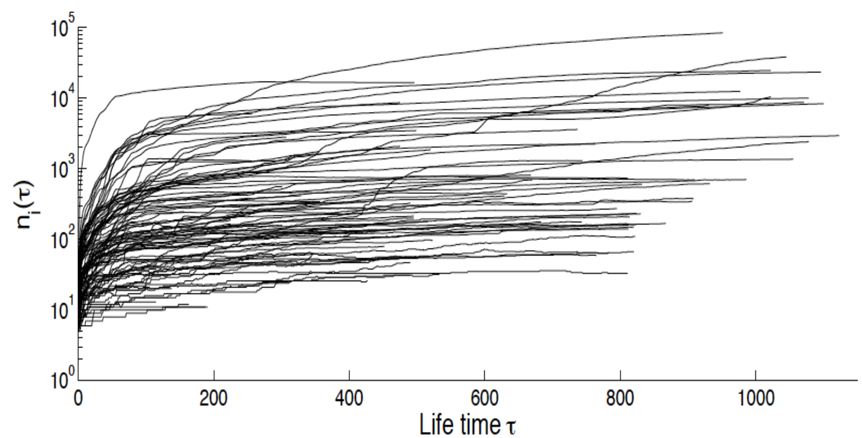
Data: Hourly data on 2,705 applications collected between 25 June 2007 and 14 August 2007

# Adoption dynamics

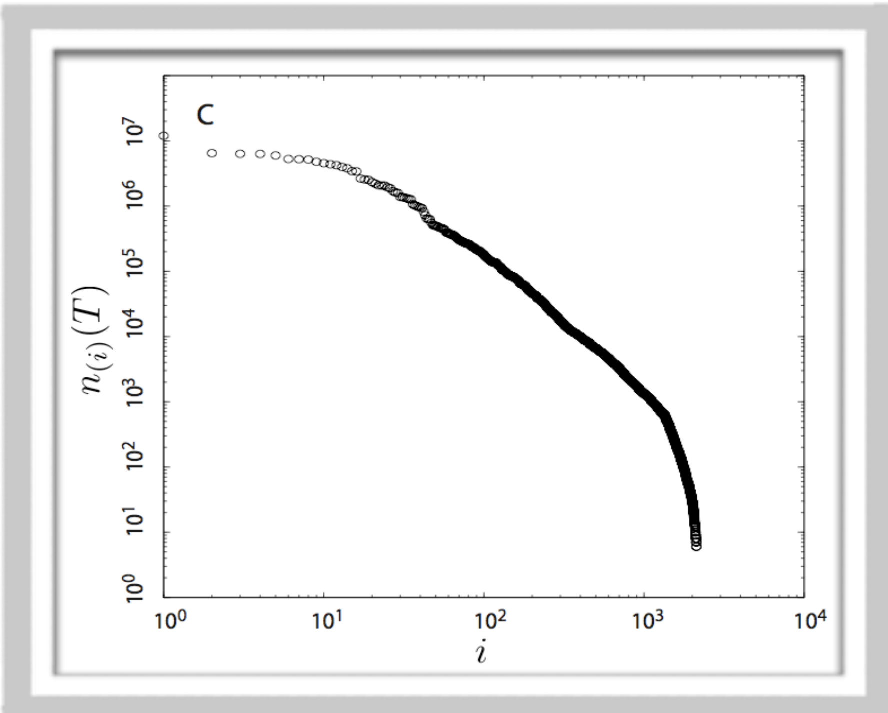


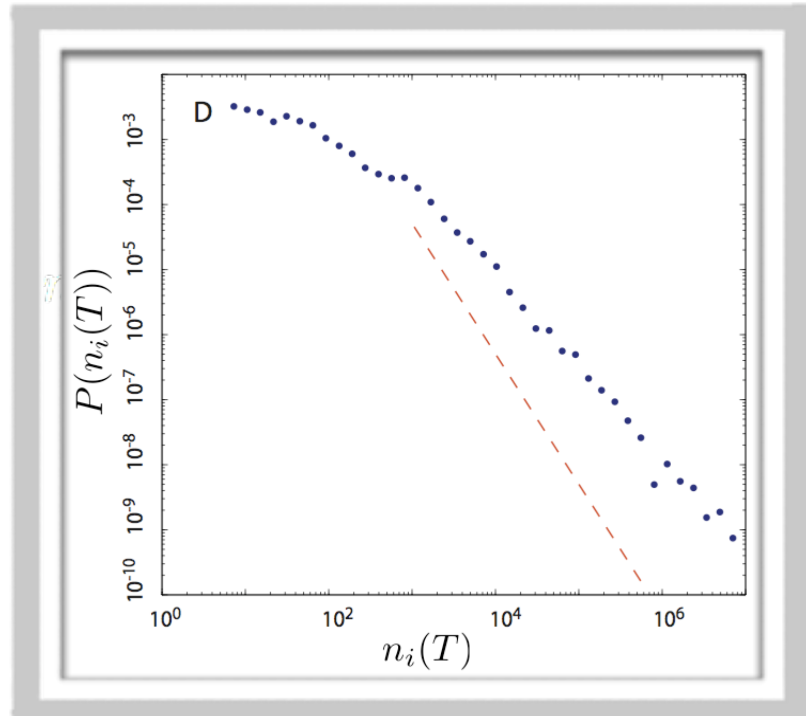



# Popularity distributions




Zipf plot




Cumulative density plot

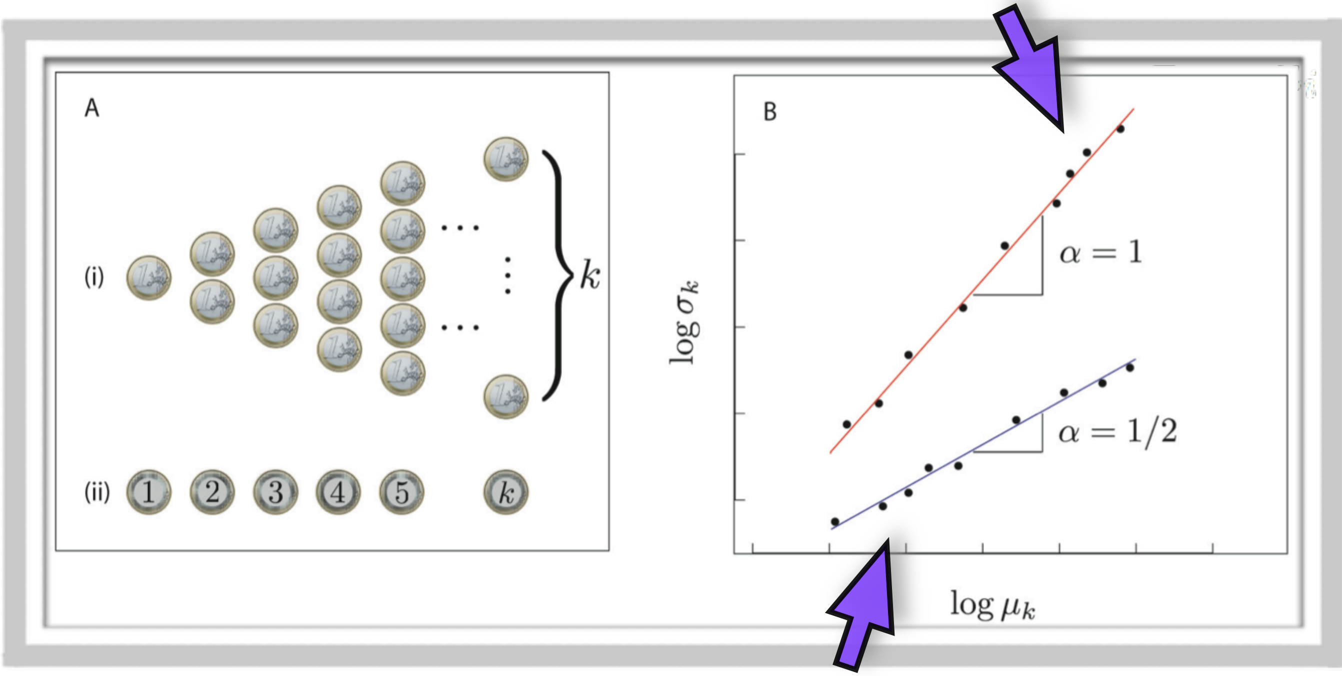
1. Top Friends (11,962,481 users)
2. Video (6,487,572 users)
3. Graffiti (6,335,873 users)
4. My Questions (6,324,224 users)
5. iLike (5,988,584 users)
6. Free Gifts (5,282,413 users)
7. X Me (5,236,443 users)
8. Superpoke! (5,175,439 users)
9. Fortune Cookie (4,774,815 users)
10. Horoscopes (4,555,010 users)

# Correlations revealed by temporal FS

CORRELATED

INDEPENDENT

CORRELATED




log-log plot

$\sigma_i \sim \mu_i^{\alpha}$

INDEPENDENT

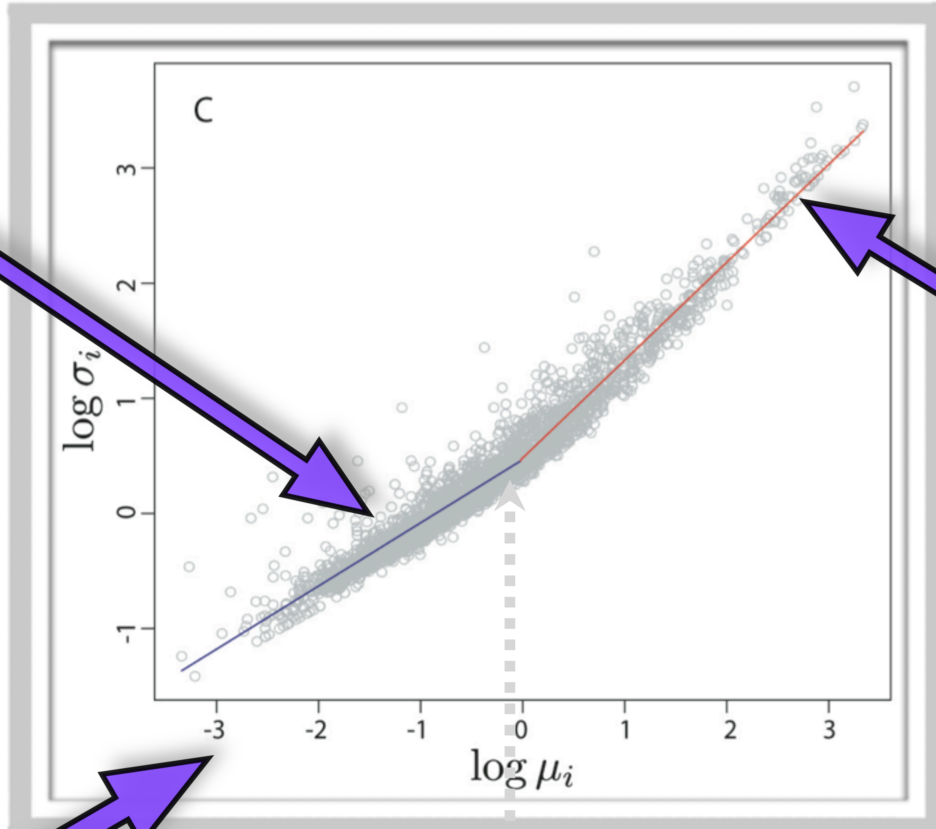
$$\mu_i \equiv \langle f_i \rangle = \frac{1}{T_i} \sum_{t=1}^{T_i} f_i(t) \qquad \sigma_i = \left( \frac{1}{T_i - 1} \sum_{t=1}^{T_i} \left[ f_i(t) - \langle f_i \rangle \right]^2 \right)^{1/2}$$

Installation of Facebook applications corresponds to having a huge set of biased heterogeneous coins, one per application for each user

"Coin tosses" are now influenced by both local and global information

# Tipping point in scaling behaviour

α≈0.55




α≈0.85

x-axis interpretation:
no. of new installations

Individual regime

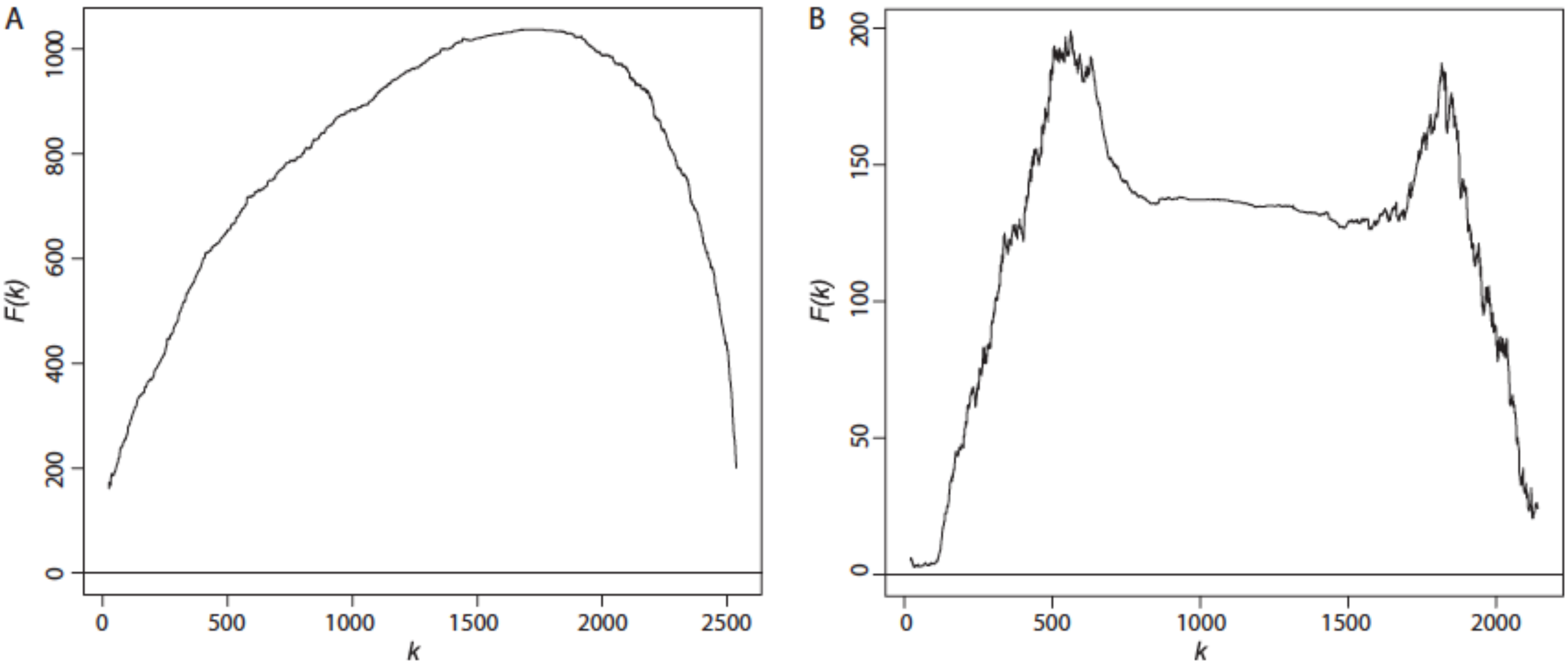
Collective regime

$\alpha_I \approx 0.55$

$\alpha_C \approx 0.85$

Breakpoint corresponds to approx. 55 installations per day

# Breakpoint analysis

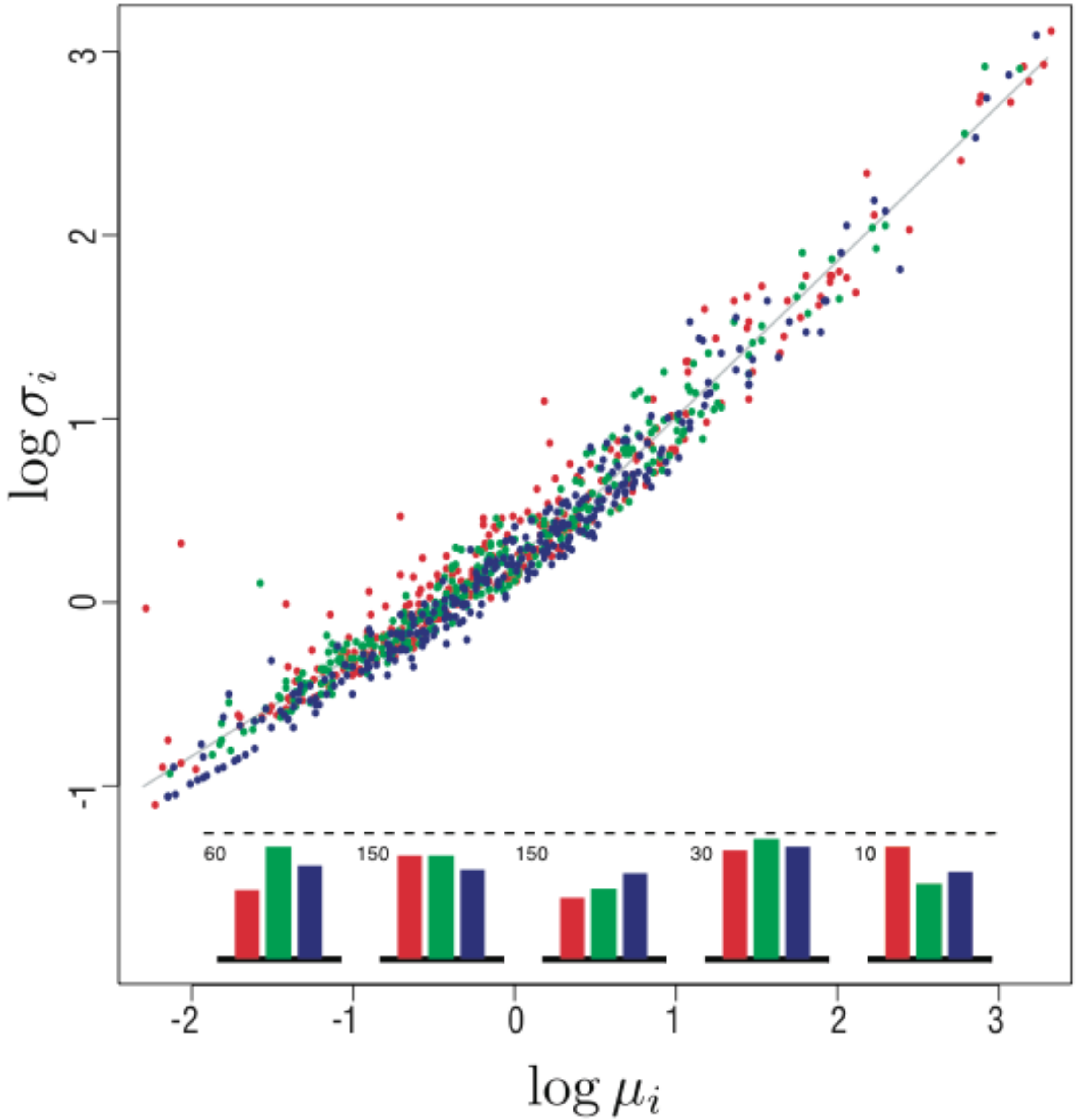



**(A)** F-statistic smooth and well-behaved. Maximum at $F_{(k)} \approx 1035$ for observation k=1795, corresponding to $\log(\mu_{(1795)}) \approx 0.36$.

**(B)** No statistical evidence for breakpoint.

Zeileis, Kleiber, Krämer and Hornik (2003). Testing and dating of structural changes in practice, *Computational Statistics and Data Analysis* **44**: 109–123.

# Effect of application lifetime on scaling



# Constructing the synthetic time series

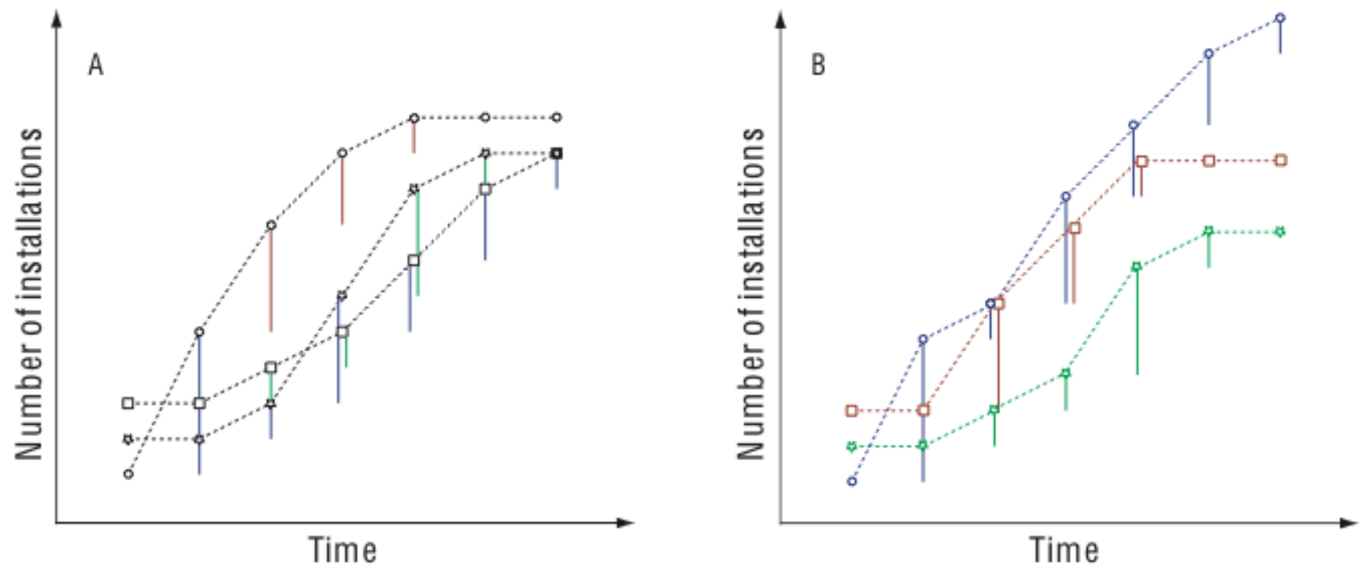
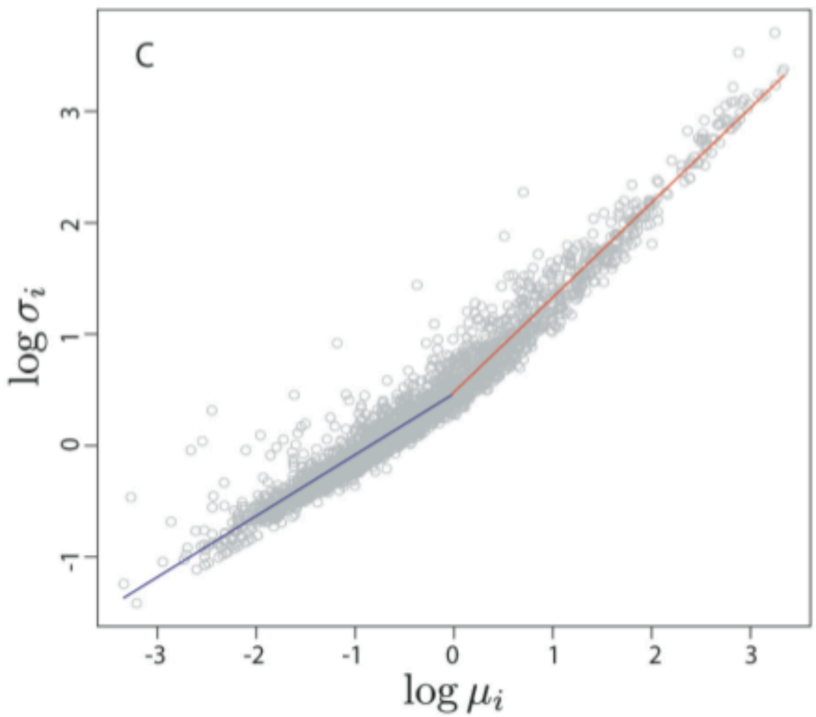
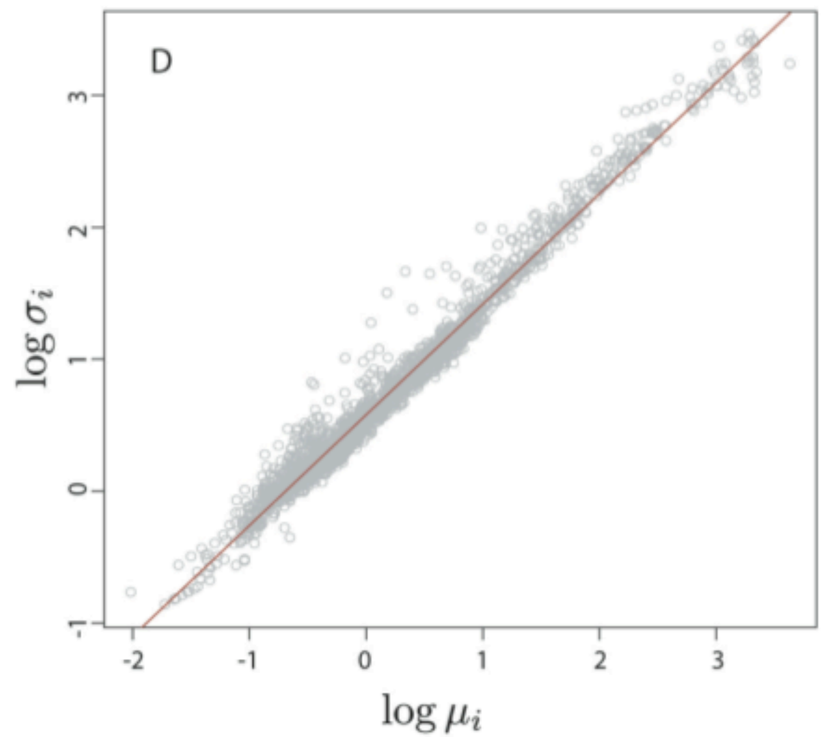



Figure 2: Schematic of the construction of the synthetic time series $\tilde{n}_i(t)$. **(A)** The empirical data consists of $t = 1, \ldots, 7$ observations for three applications. The data points have been connected with dashed black lines to guide the eye. For the most popular application at time $t-1$, the change in number of users between times $t-1$ and $t$ is indicated by the height of the vertical red bar at time $t$, which corresponds to $\tilde{f}_1(t)$ in the text. Similarly, $\tilde{f}_2(t)$ and $\tilde{f}_3(t)$ are indicated by the green and blue bars, respectively. An easy way to understand the process is first to compute the difference in the number of users for all applications given by $f_i(t) = n_i(t) - n_i(t-1)$ and then color the difference based on $r_i(t-1)$, the rank of the application at time $t-1$. **(B)** The synthetic time series are seeded by the initial values taken from the empirical data such that $\tilde{n}_1(1) = n_\square(1)$, $\tilde{n}_2(1) = n_\star(1)$, and $\tilde{n}_3(1) = n_\circ(1)$ of the empirical data and they are constructed by adding together the difference bars of the same color. Overlapping bars have been shifted slightly horizontally for clarity of presentation.

# Empirical vs. synthetic data




This is a key comparison since we are restricted to observational data.

# Interim conclusions

In the Facebook environment we are able to track actions relating to all users and applications, rather than a subset of both. Importantly, for the period in which data were collected, exogenous drivers (e.g. media campaigns) can also be largely excluded. This provides an unusually clean and complete setting in which to study innovation diffusion. Of course, we are restricted to observational data, and cannot trace the underlying network structure.

The two distinct regimes that we observe are novel. Also, note the difference with standard epidemic spread models, where there is no global signal.

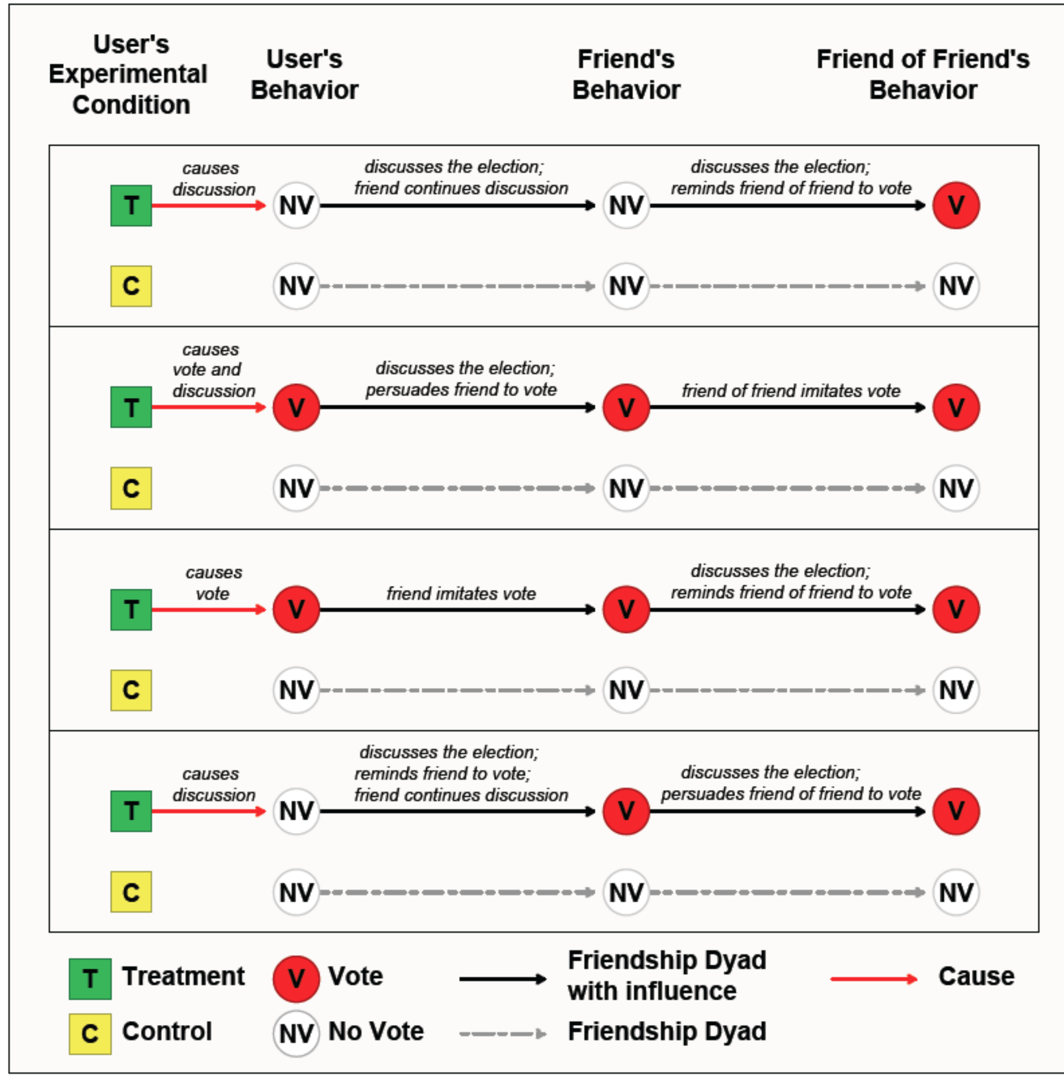
Key open question. Can we use purely observational data to differentiate between different potential mechanisms such as social influence?

Onnela & FRT (2010). Spontaneous emergence of social influence in online systems, *Proceedings of the National Academy of Sciences* **107**: 18375-18380.

# An online experiment in social influence

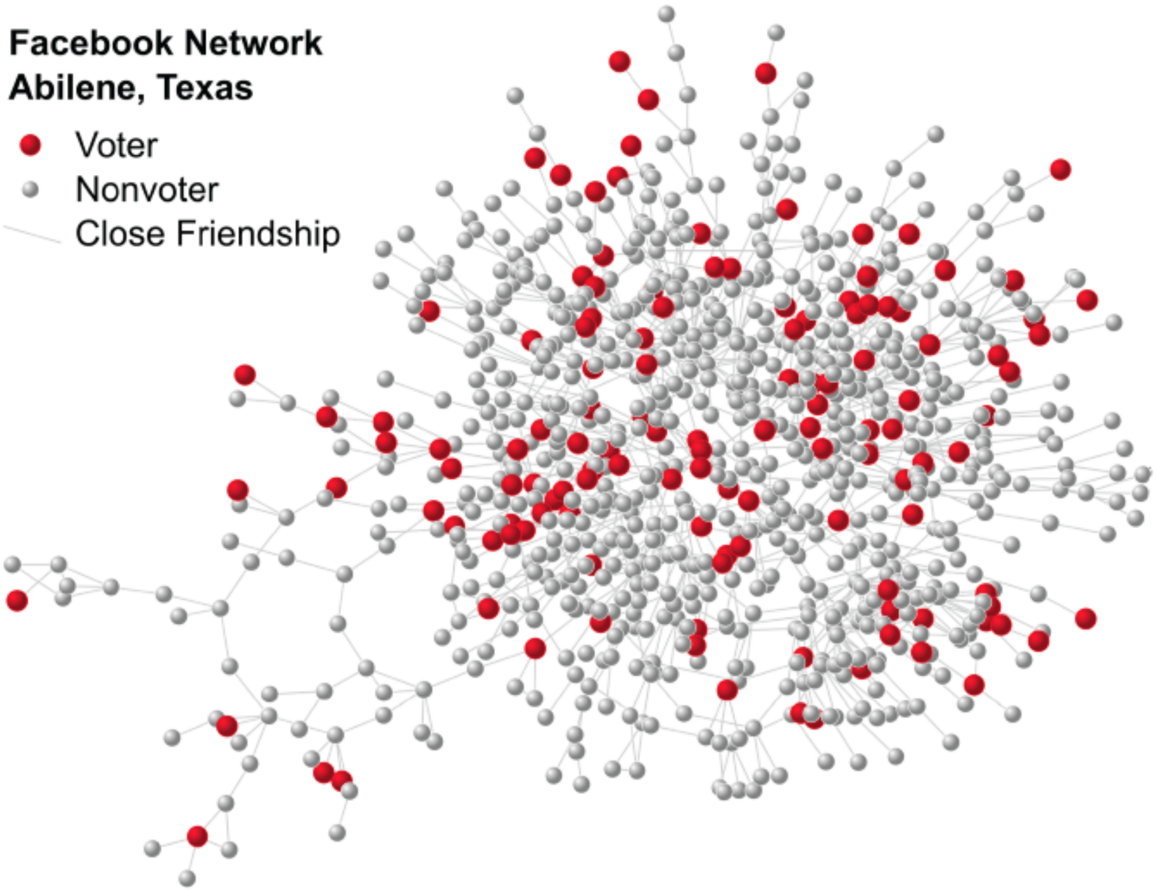
## Informational Message

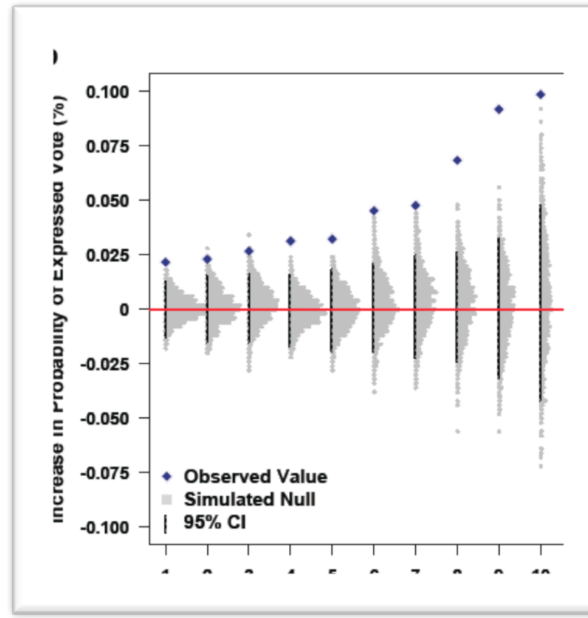

## Social Message


R M Bond et al. (2012). A 61-million-person experiment in social influence and political mobilisation, *Nature* **489**:295-298.

# Voter mobilisation on Facebook



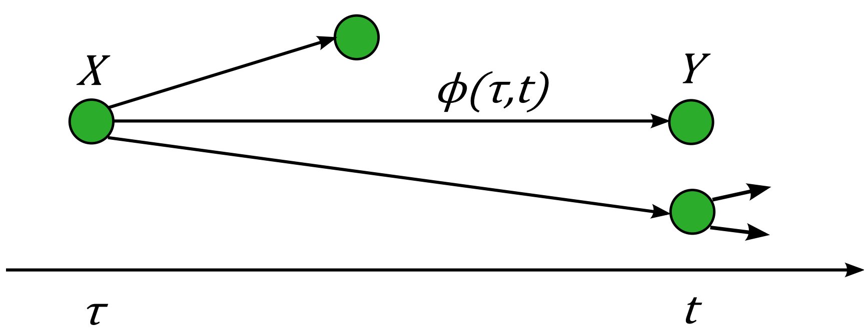

R M Bond et al. (2012). A 61-million-person experiment in social influence and political mobilisation, *Nature* **489**:295-298.

# Possible (simple) generative models

Preferential attachment (cumulative advantage)

Random-copying model (imitate recent choices of others)

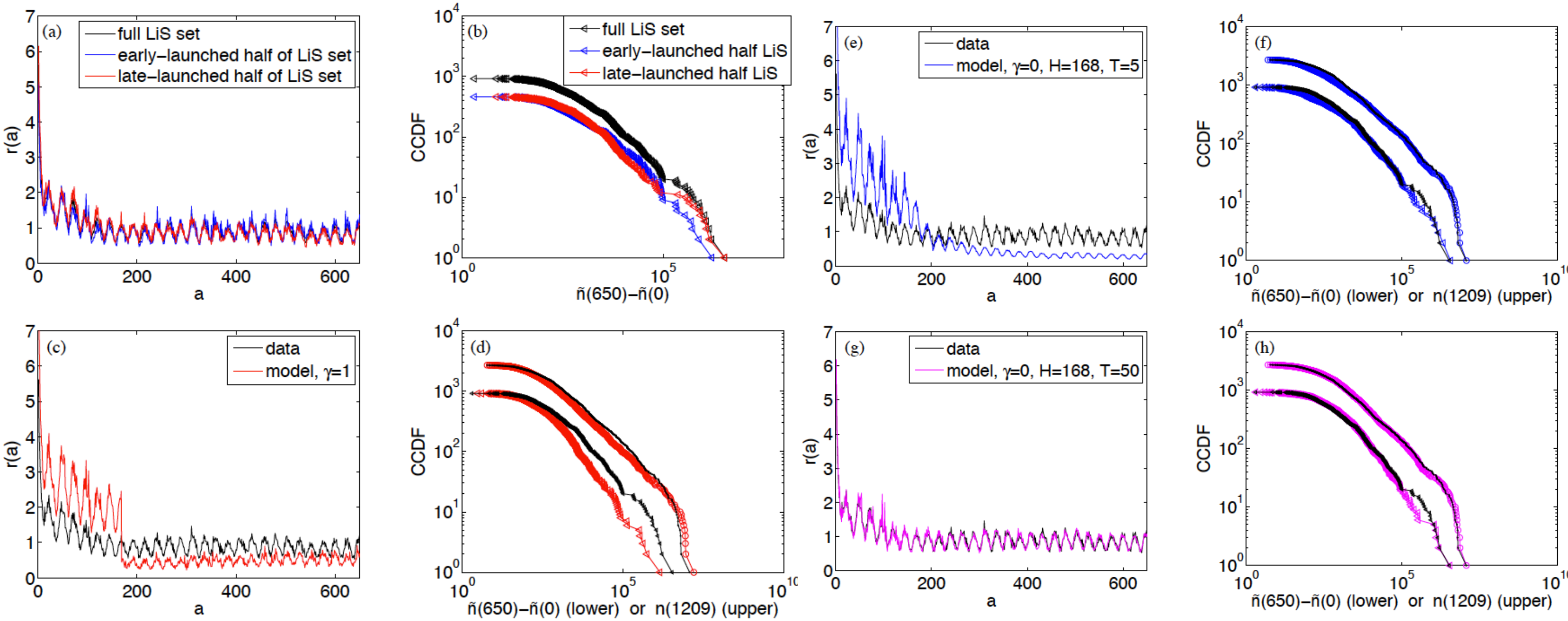
Generalized random-copying model:




H - history window
T - response time parameter
W( $\tau$ ) - memory weighting function
$\gamma$ - fraction of installs using cumulative information

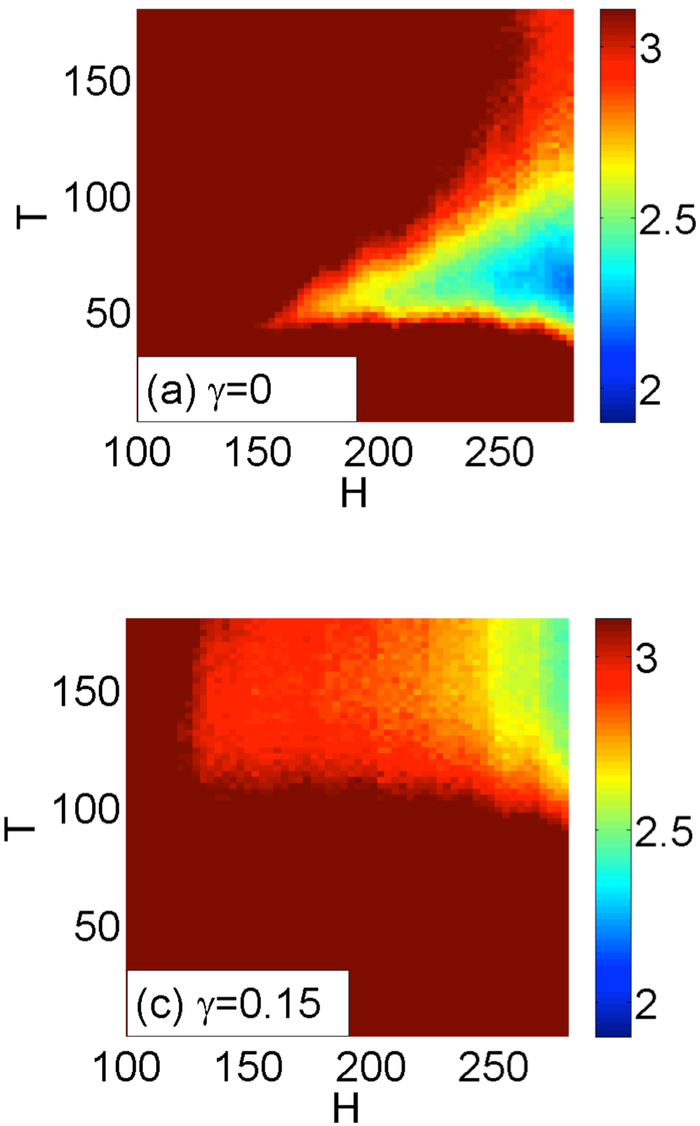
J P Gleeson, D Cellai, J.-P. Onnela, M A Porter, FRT (2013), A simple generative model of collective online behaviour, *arXiv*:1305.7440.

# Comparing models and data



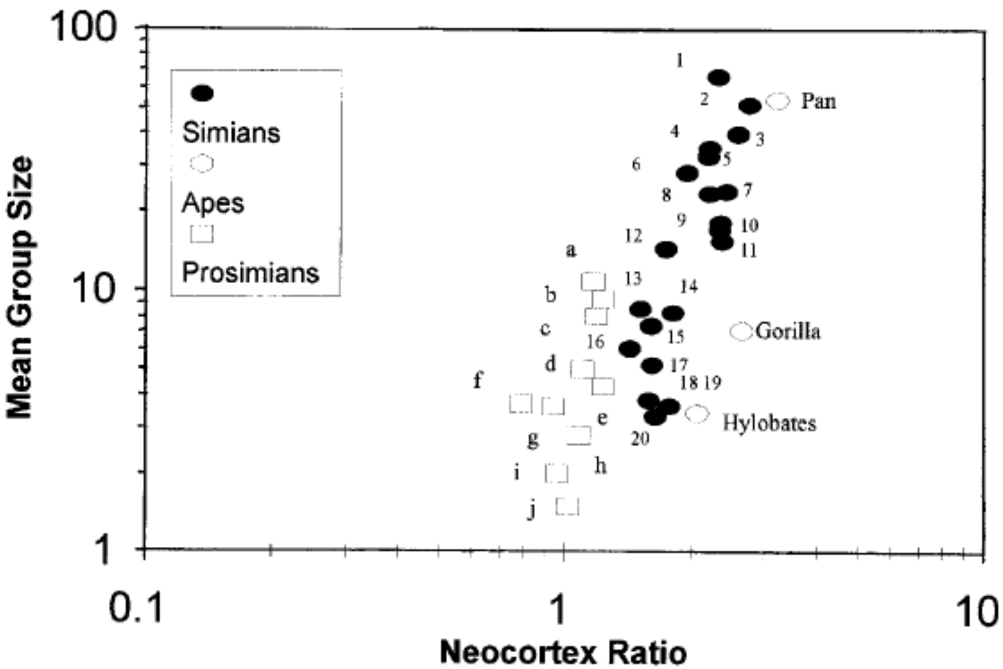

a – age
r(a) – mean scaled growth rate
CDDF – complementary cumulative distribution function

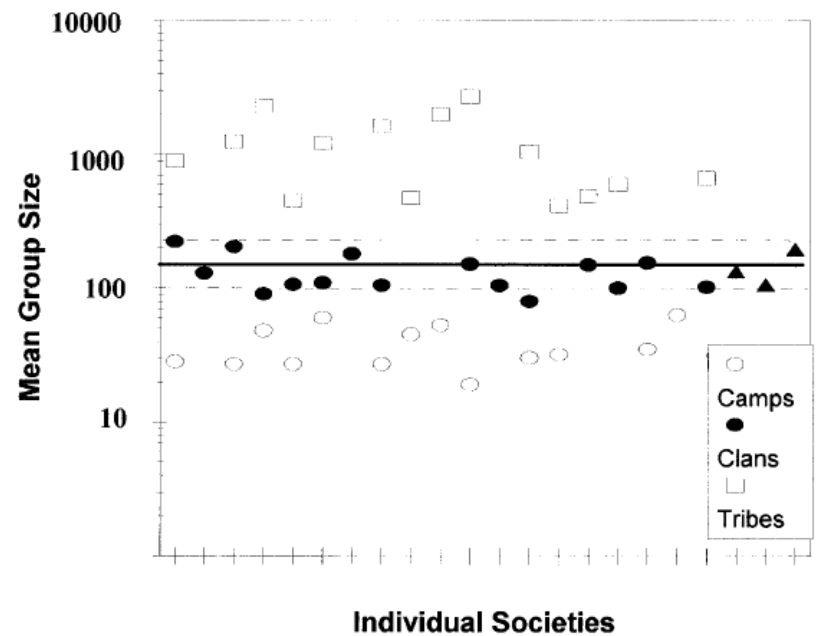
# Fitting the model to temporal data




Parameter planes for different values of $\gamma$ showing the $L^2$ norm of the difference between the simulated *r(a)* curve from the recent-information-dominated model and the data.

# The social brain hypothesis

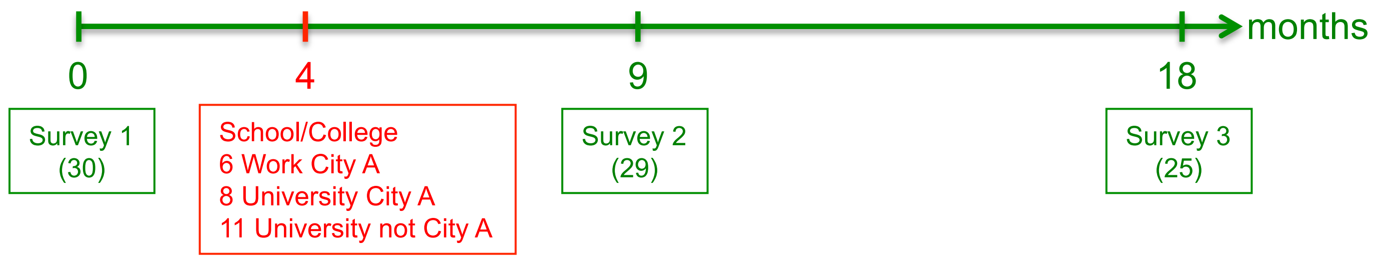



"Dunbar number" ≈ 150 individuals

R I M Dunbar (1998), The social brain hypothesis, *Evolutionary Anthropology*, **6**: 178-190.

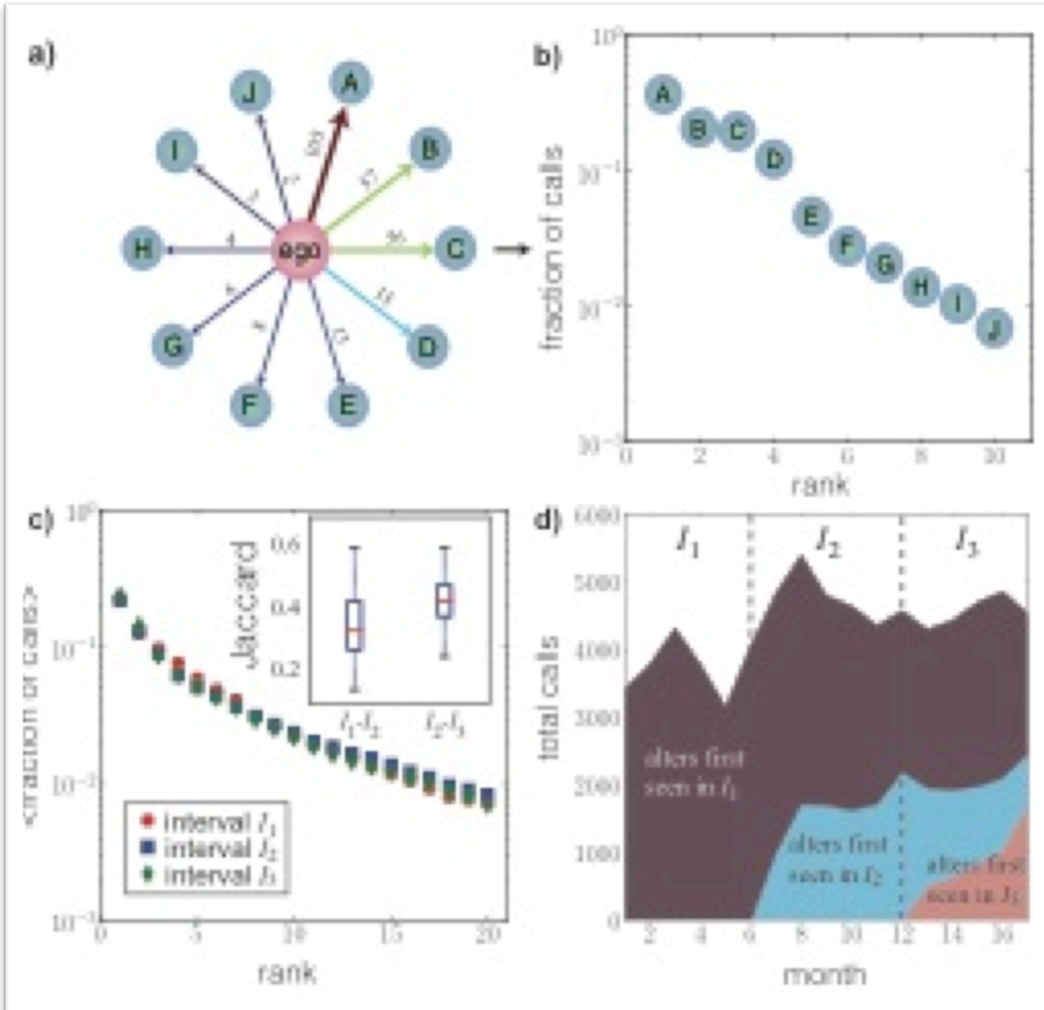
# Dynamics of kinship and friendship ties




- 30 students in City A recruited in final year of secondary school. 25 completed all parts of the study.
- Provided 18-month cellphone plan with free calls and texts
- 3 surveys recording kinship and friendship ties, shared activities, time to last contact, emotional closeness, etc.
- At month 4 students finish school, and move to university either in City A or elsewhere, or start work in City A.
- Participants asked to list all known and living relatives (kin), and all friend and acquaintances (1217 network members – 479 kin, 738 friends).

S G B Roberts & R I M Dunbar (2011), The costs of family and friends: an 18-month longitudinal study of Relationship maintenance and decay, *Evolution and Human Behavior*, **32**: 186-197.

# Social signatures and network turnover




(a) The number of calls to each alter is counted.

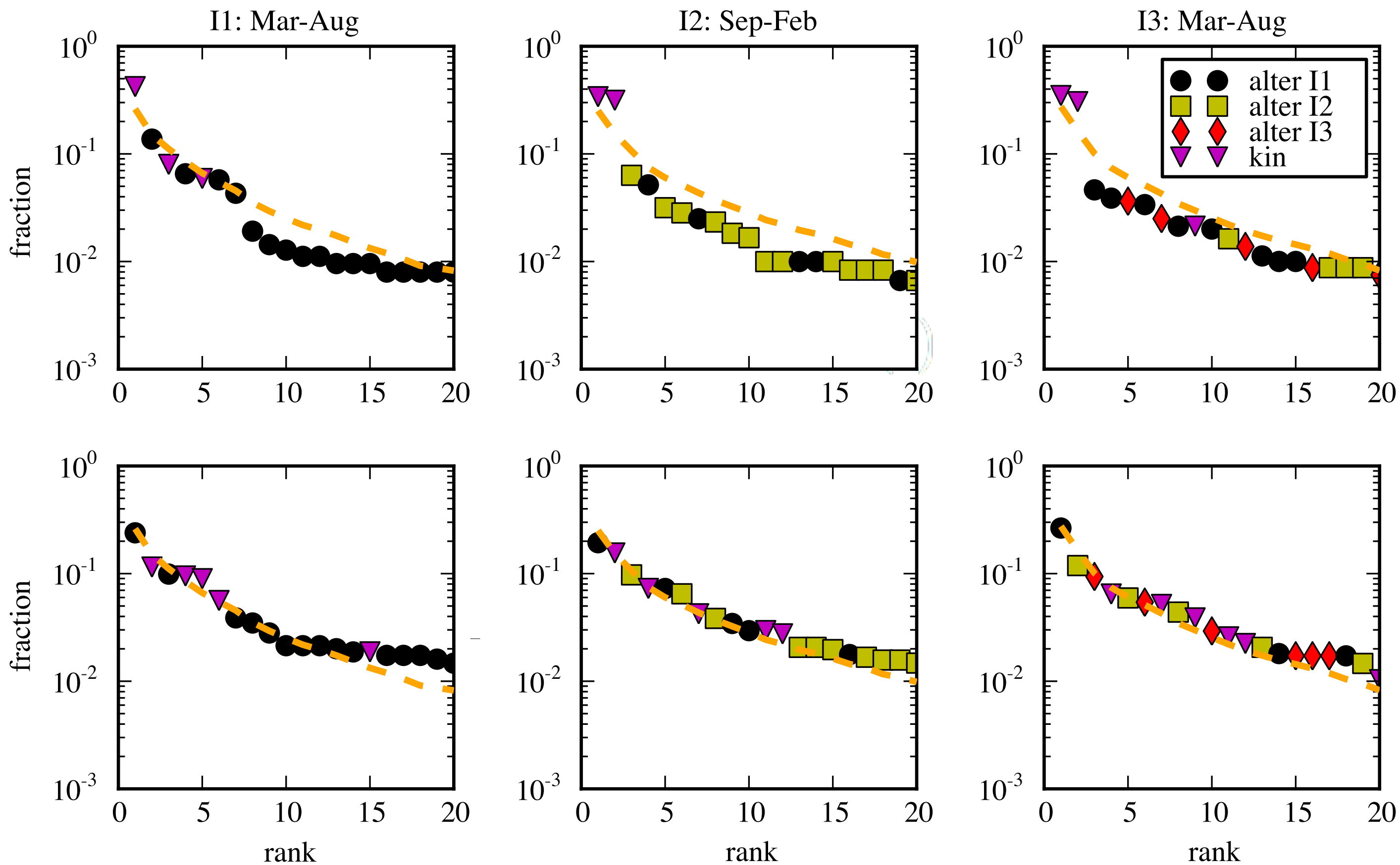
(b) Social signatures are constructed by ranking the number of calls for each ego, and then calculating the fraction of calls to the alter of each rank.

(c) Averaging the social signatures over the set of participants for three consecutive 6-month intervals, their shape is invariant as indicated by the Jaccard index between 20 top-ranking alters (inset).

(4) Network turnover is significant.

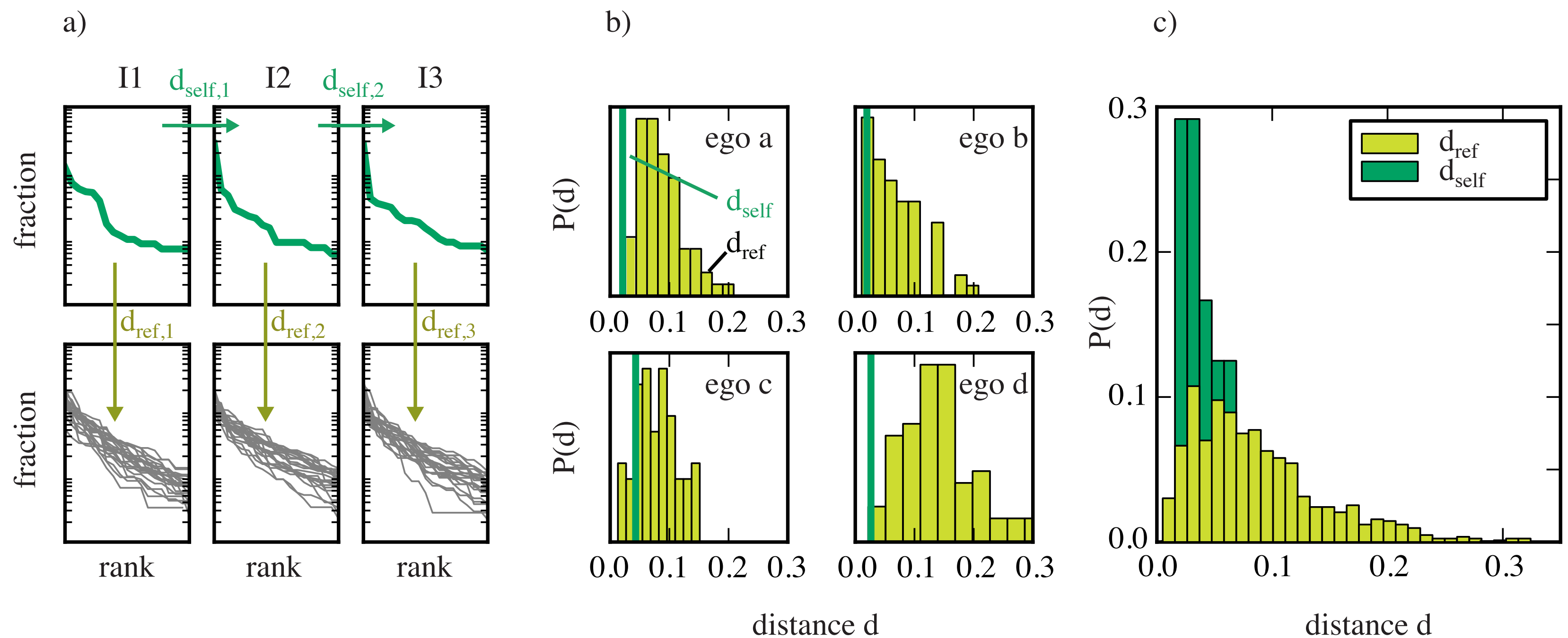
J Saramäki, E A Leicht, E López, S G B Roberts, FRT, R I M Dunbar (2012), The persistence of social signatures in human communication, *arXiv*:1204.5602.

# Evolution of social signatures




The top row depicts the social signature of a male participant who went to university in another city, and the bottom row represents a female who went to university in City A. The symbols correspond to alters observed for the first time in interval $I_1$ (circles), $I_2$ (squares), and $I_3$ (diamonds), or to kin (triangles) as reported by the egos. The dashed line indicates the social signature averaged over all 24 egos.

# Persistence of individual social signatures

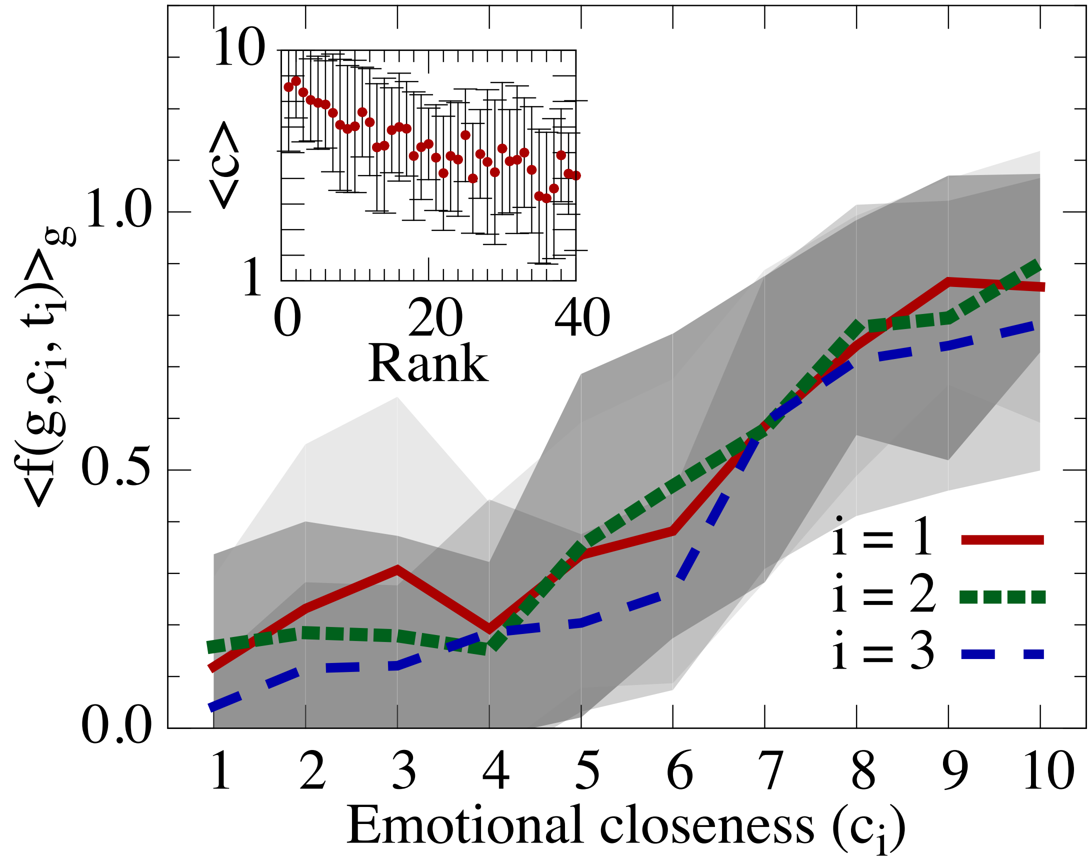



a) Distances between social signatures based on Jensen-Shannon divergences. For the focal ego (top row) self-distances $d_{self}$ are calculated for consecutive intervals and averaged. Reference distances $d_{ref}$ are calculated for each interval between the social signatures of the focal ego and all other egos (bottom row).

b) Average value of $d_{self}$ and histograms of $d_{ref}$ for four sample egos.

c) Distributions of $d_{self}$ and $d_{ref}$ for all egos.

# Call patterns and emotional closeness




Fraction of alters, averaged over all egos, that are actually called by their ego in a 6-month period, $I_1$, given the ego scores the alter with emotional closeness $c_i$. The shaded region indicates standard deviation. The inset shows the average emotional closeness of alters of varying rank with error bars showing the standard deviation. The inset shows the average emotional closeness of alters of varying rank with error bars showing standard deviations.

# Closing observations

Our 18-month longitudinal dataset gives us an interesting opportunity to try to combine traditional survey-based data with time-stamped call data.

Our initial findings suggest considerable heterogeneity among individuals as to how they allocate time to their social relationships, but stability of individual patterns over time. This applies even when alters change.

Clear limitation is the size of the population being studied. Hence, question whether we can use other datasets to validate what we see.

More general question is how one can link small-scale high but very rich datasets with large datasets that include only rudimentary information on individuals.

# Acknowledgements


<u>Collaborators:</u>

Davide Cellai (Limerick)
Robin Dunbar (Oxford)
James Gleeson (Limerick)
Elizabeth Leicht (Oxford)
Eduardo López (Oxford)
Jukka-Pekka Onnela (Harvard)
Mason Porter (Oxford)
Sam Roberts (Chester)
Jari Saramäki (Helsinki)

EPSRC
Engineering and Physical Sciences Research Council

SEVENTH FRAMEWORK PROGRAMME

ICT collective