http://www.cabdyn.ox.ac.uk/

Innovation adoption
in an online social network

Felix Reed-Tsochas
CABDyN Complexity Centre
University of Oxford

CCSS Colloquium
ETH Zürich, 5 April 2011

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
- Microscopic models
- Extension to other online datasets

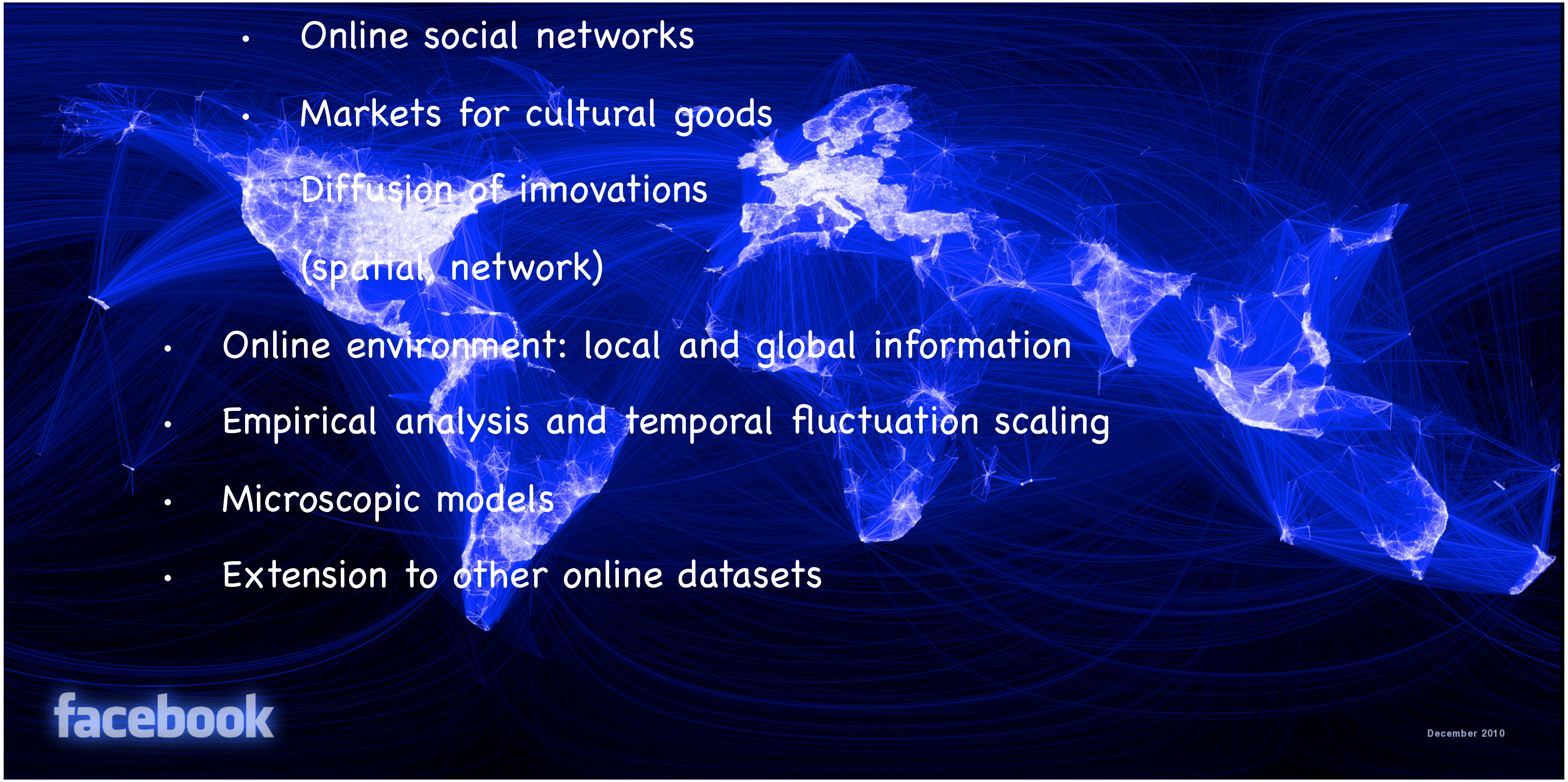
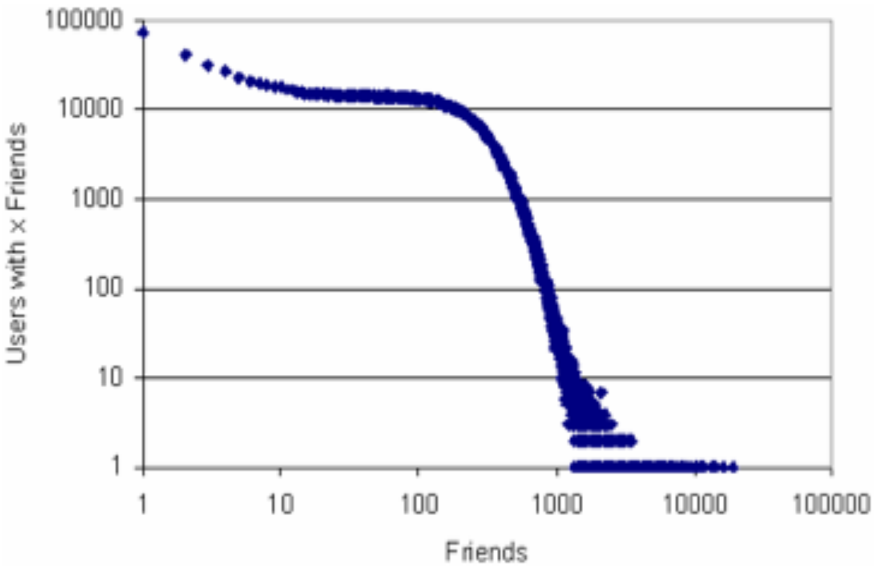
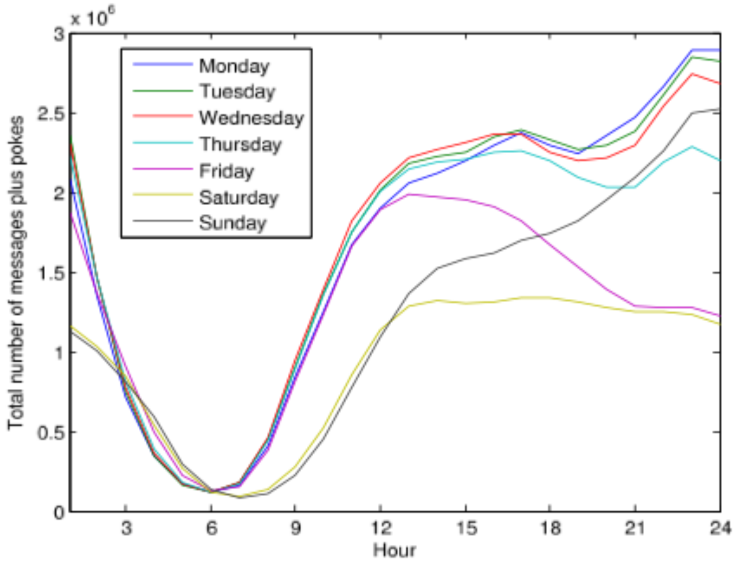


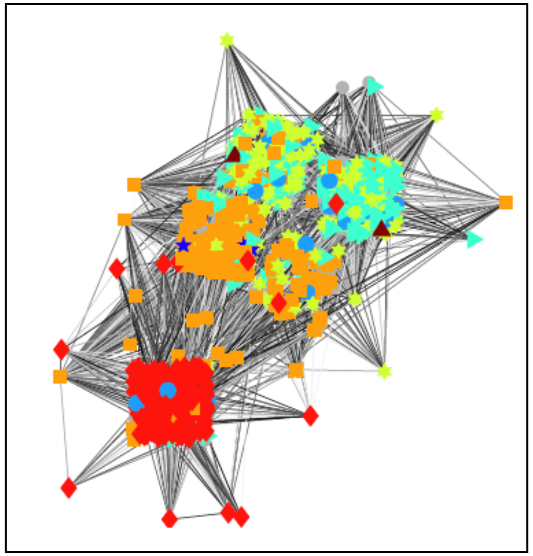
Image by Paul Butler, see http://www.facebook.com/notes/facebook-engineering/visualizing-friendships/469716398919

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 (2011)
Social structure of Facebook networks
arXiv:1102.2166

Cultural markets

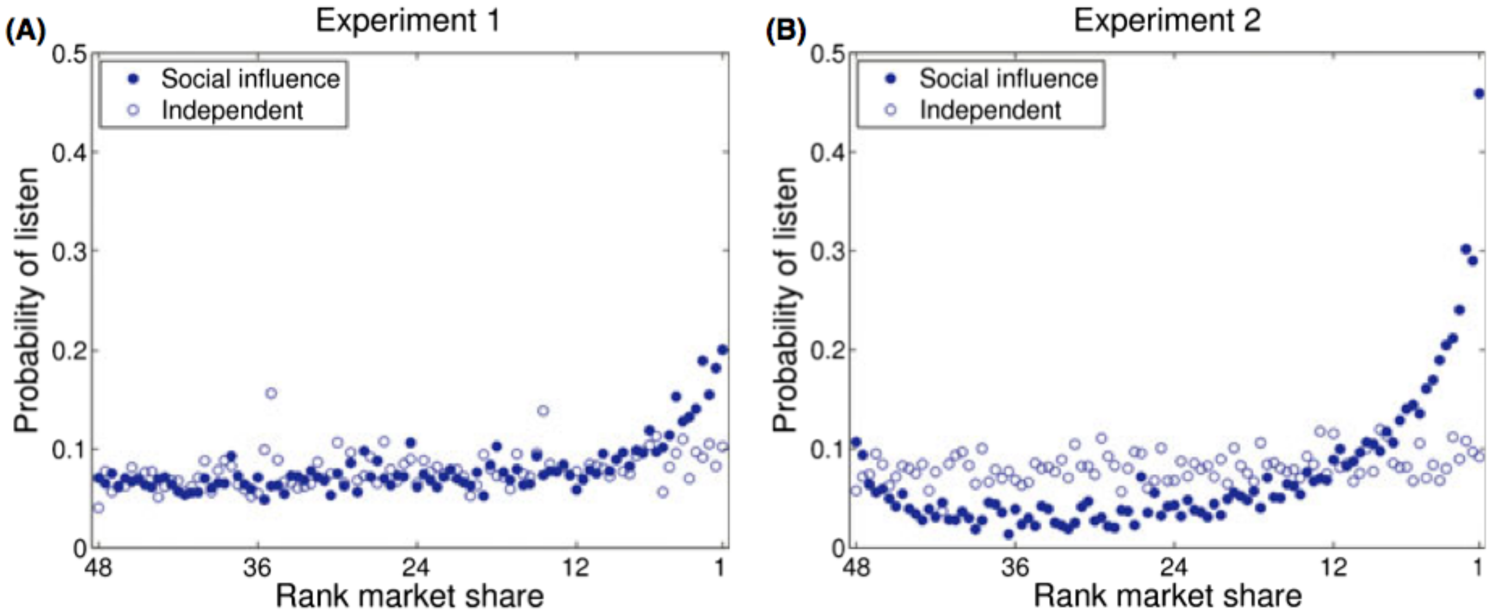
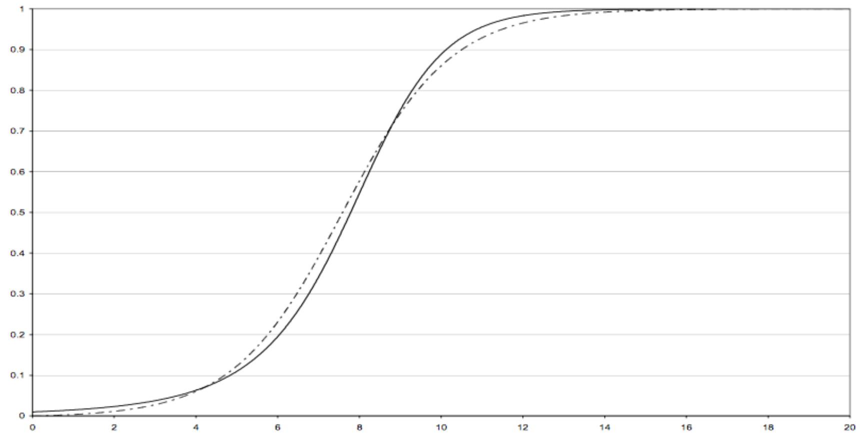


Fig. 3. Probability of listening to a song of a given market rank in Experiment 1(A) and Experiment 2(B). Participants in Experiment 2 were more likely to listen to more popular songs.

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

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

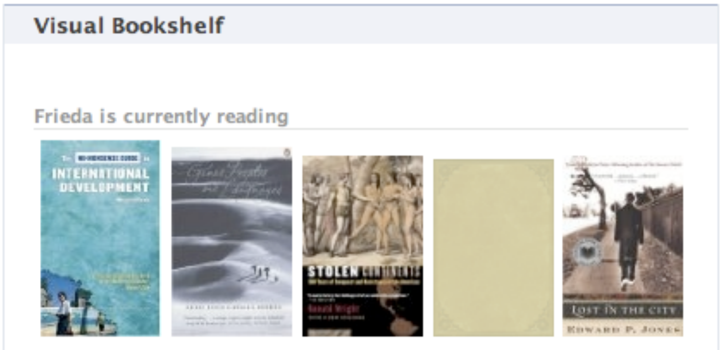
Recent re-analyses suggest that effect of sales reps etc. has been neglected

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

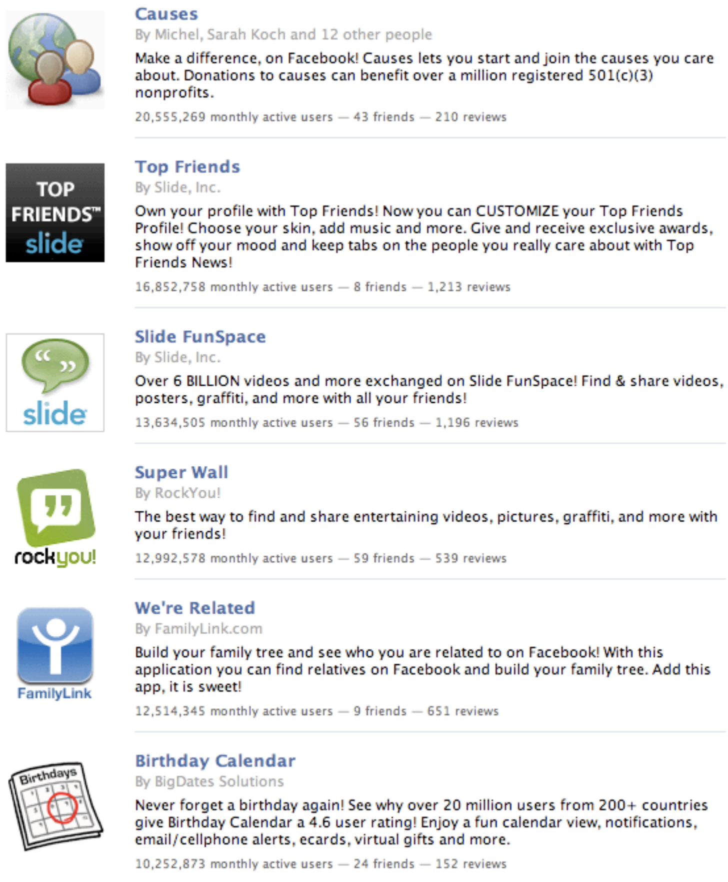
Young (2009), Innovation diffusion in heterogeneous populations: contagion, social influence, and social learning, *American Economic Review* **99**: 1899-1924.

The Facebook environment

Local information



Global information



[Note: These are from the current 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)

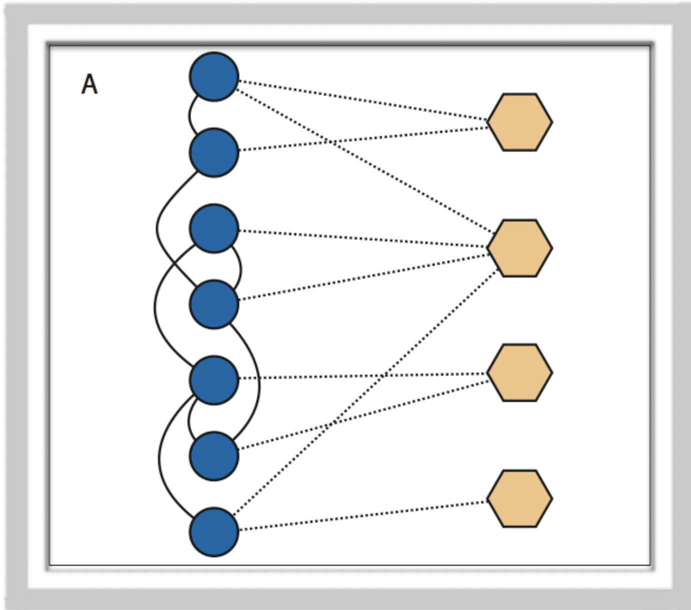
Global popularity – A high ranking may:

(i) lower search costs

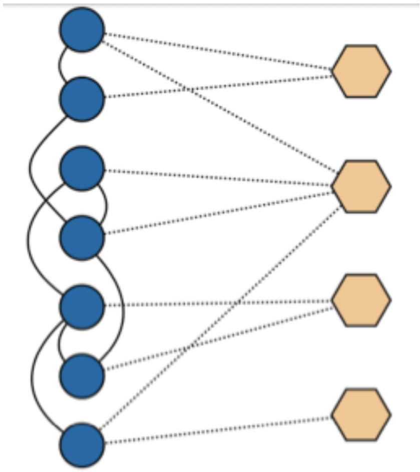
(ii) signal high quality

(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.

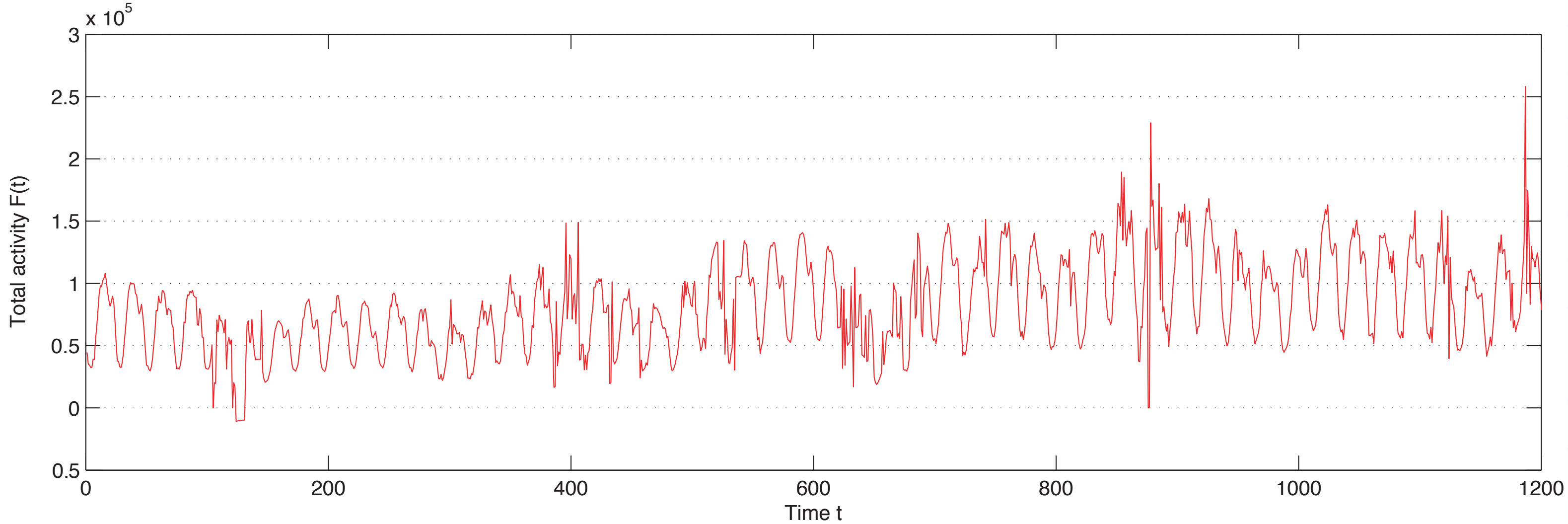
On any time scale the variance can be obtained as a time average:

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

If f is positive and additive we frequently observe:

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

Time-series activity data



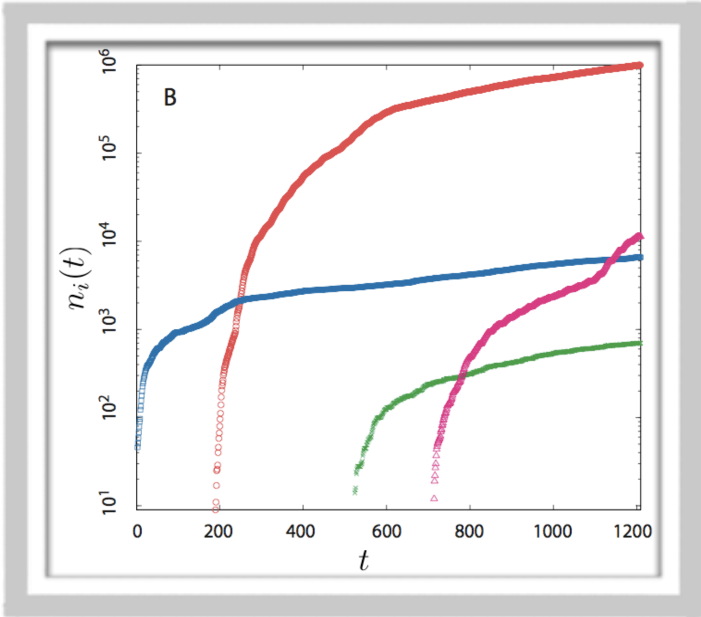
$$f_i(t) = n_i(t) - n_i(t-1) = \sum_{j=1}^{N} S_{i,j}(t)$$

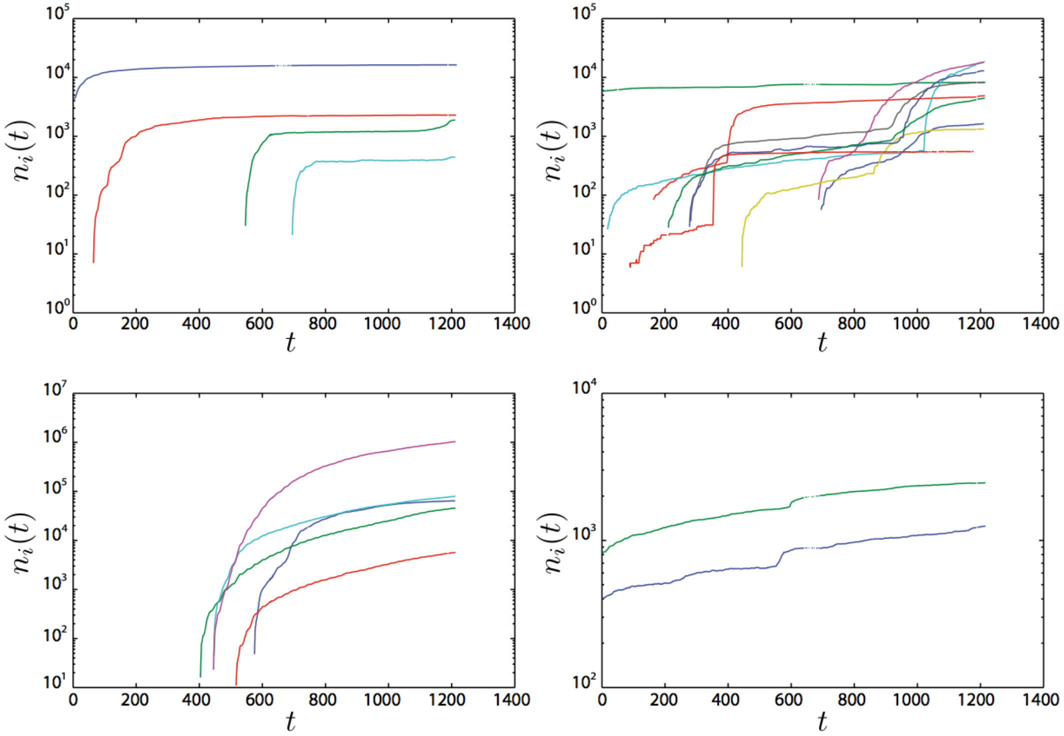
"Movies" application

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

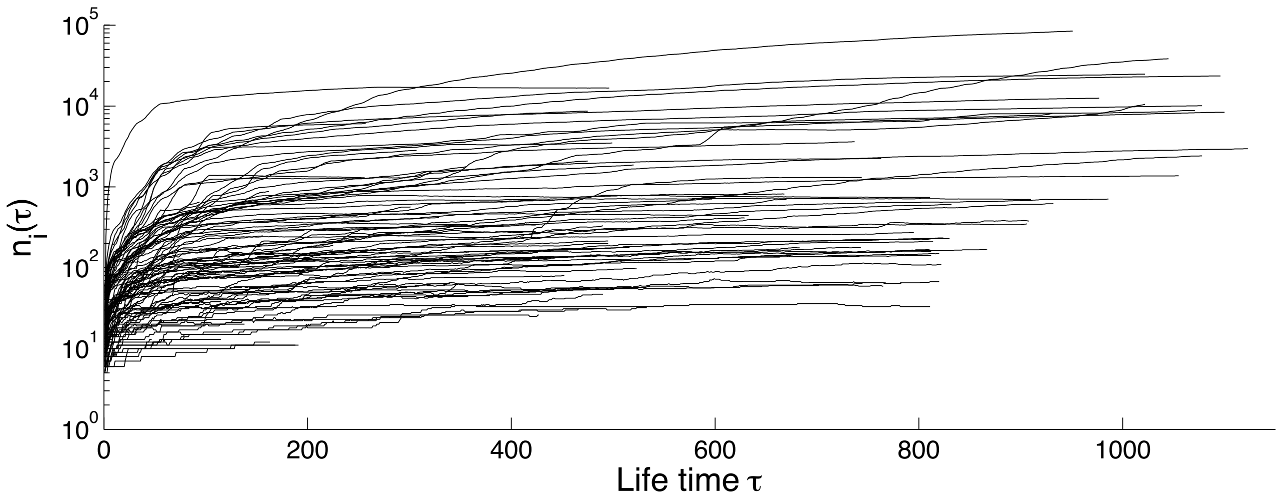
Data: Hourly data on 2,705 applications collected between 25 June 2007 and 14 August 2007

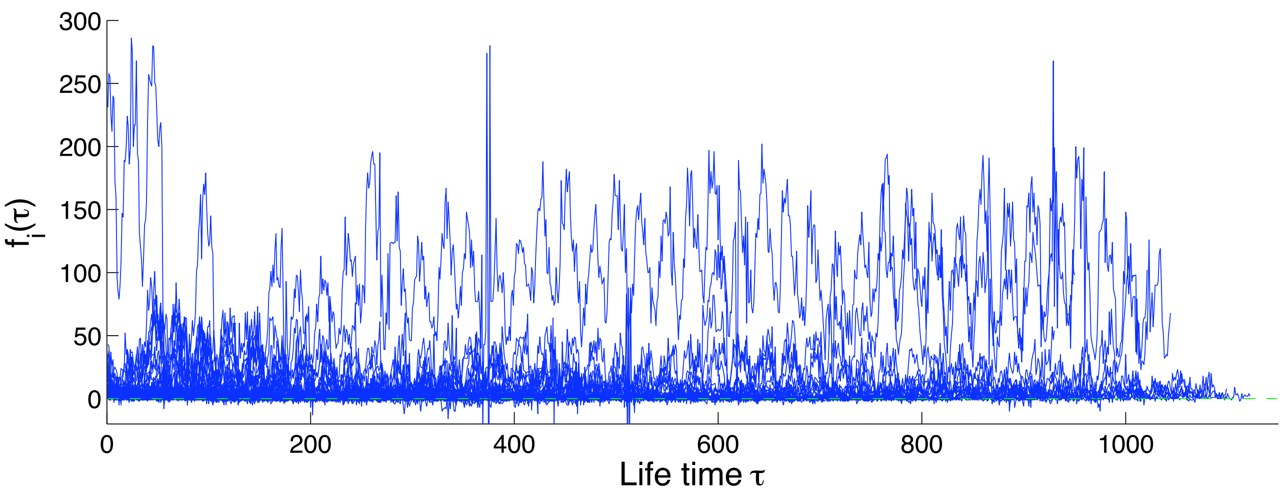
Cumulative adoption curves



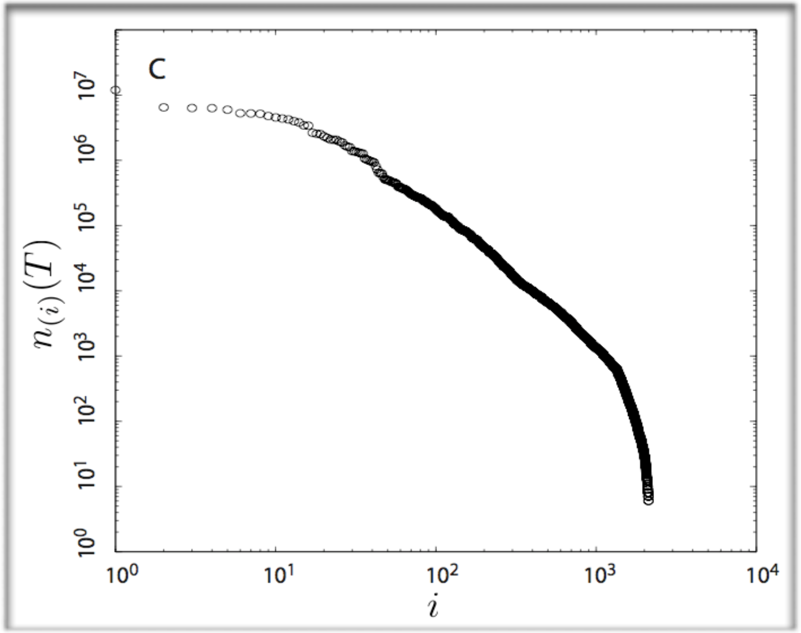


Time-shifted activity and growth

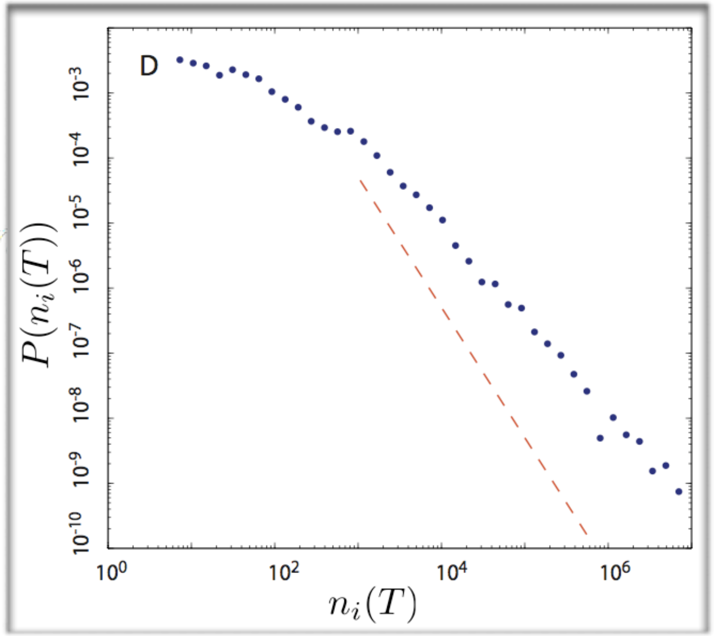




Distribution of popularity



Zipf plot



Cumulative density plot

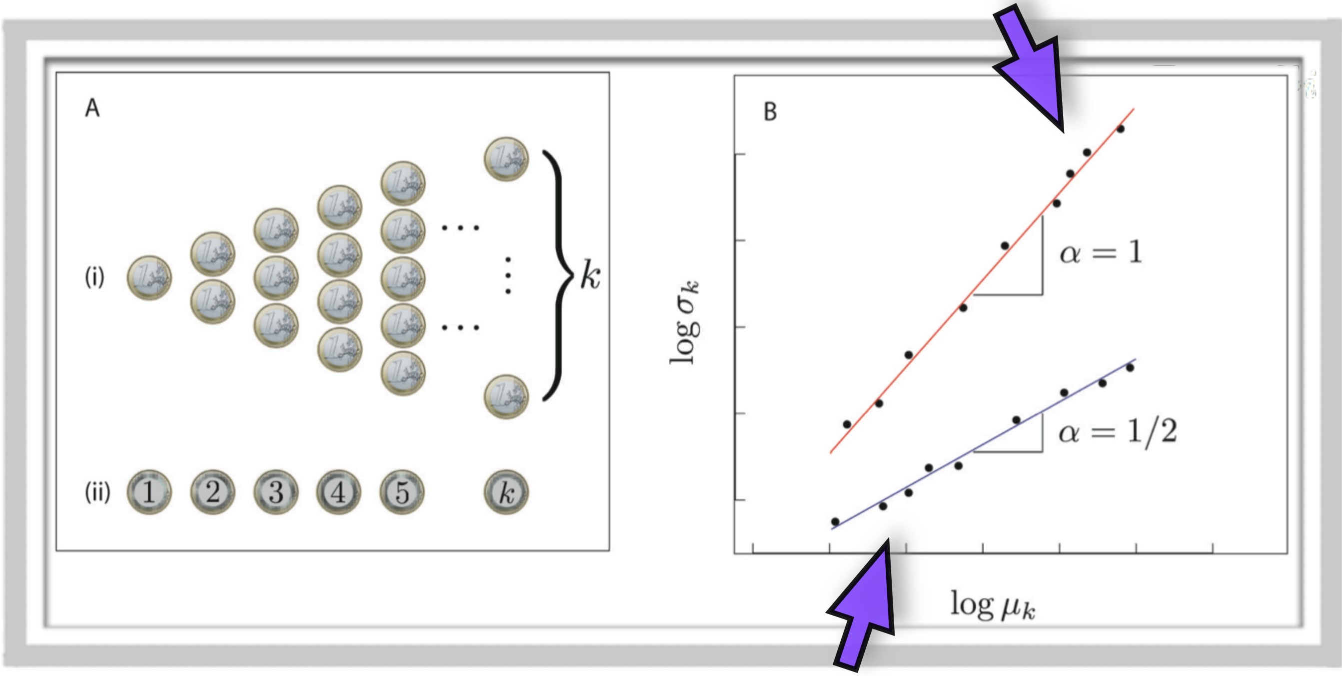
1. Top Friends (11,962,481 users)
2. Video (6,487,572 users)
3. Graffiti (6,335,873 users)
4. My Questions (6,324,224 users)
5. iLike (5,988,584 users)
6. Free Gifts (5,282,413 users)
7. X Me (5,236,443 users)
8. Superpoke! (5,175,439 users)
9. Fortune Cookie (4,774,815 users)
10. Horoscopes (4,555,010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

Correlations revealed by temporal FS

CORRELATED

INDEPENDENT

CORRELATED



INDEPENDENT

log-log plot

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

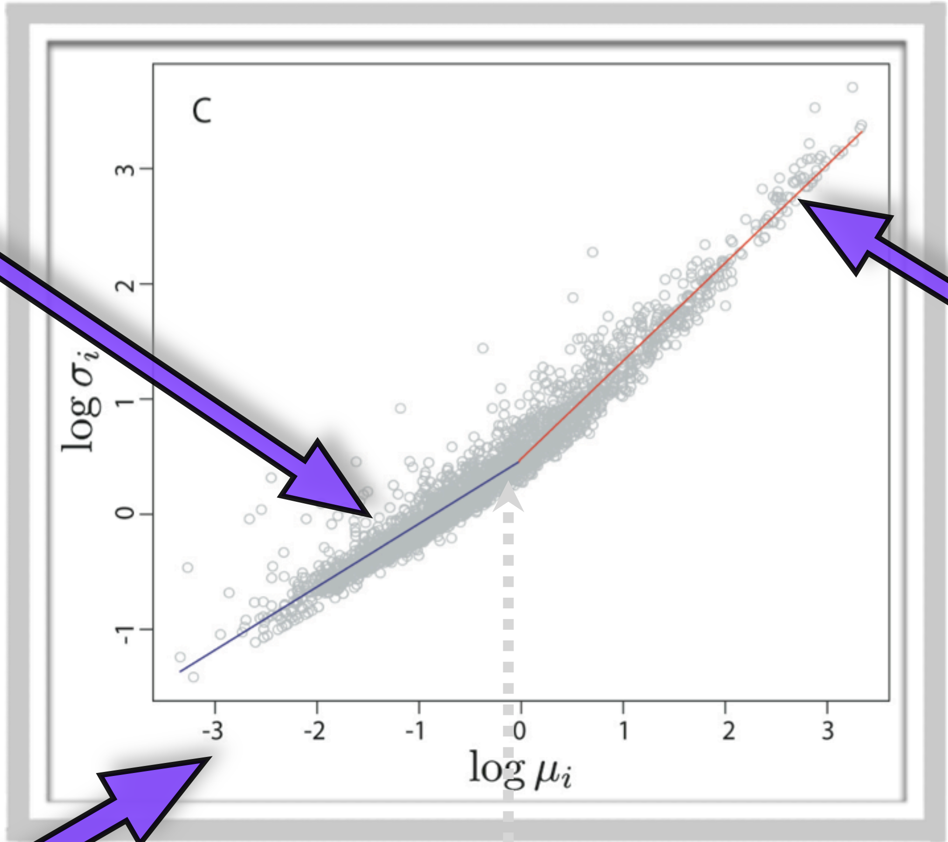
$$\mu_i \equiv \langle f_i \rangle = \frac{1}{T_i} \sum_{t=1}^{T_i} f_i(t) \qquad \sigma_i = \left(\frac{1}{T_i - 1} \sum_{t=1}^{T_i} \left[f_i(t) - \langle f_i \rangle \right]^2 \right)^{1/2}$$

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

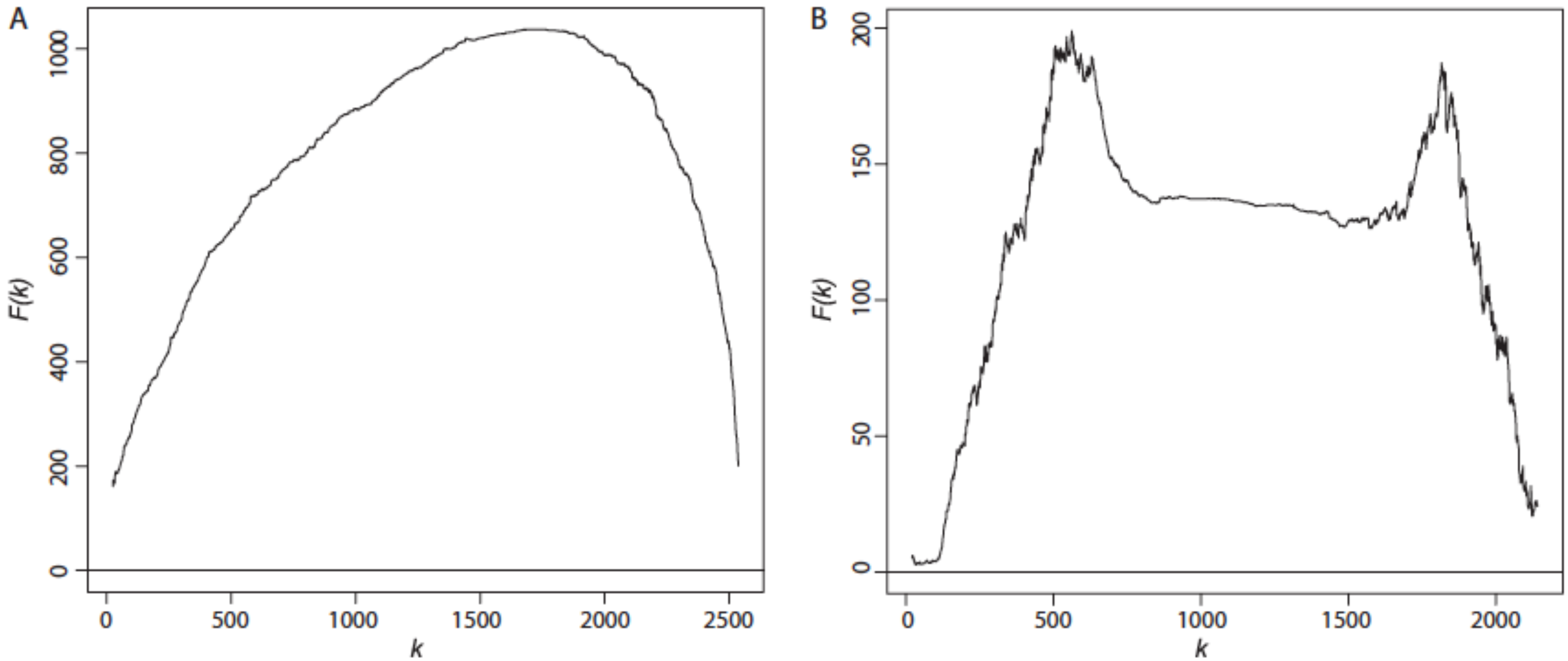
x-axis interpretation:
no. of new installations

Individual regime

Collective regime

Breakpoint corresponds to approx. 55 installations per day

Breakpoint analysis

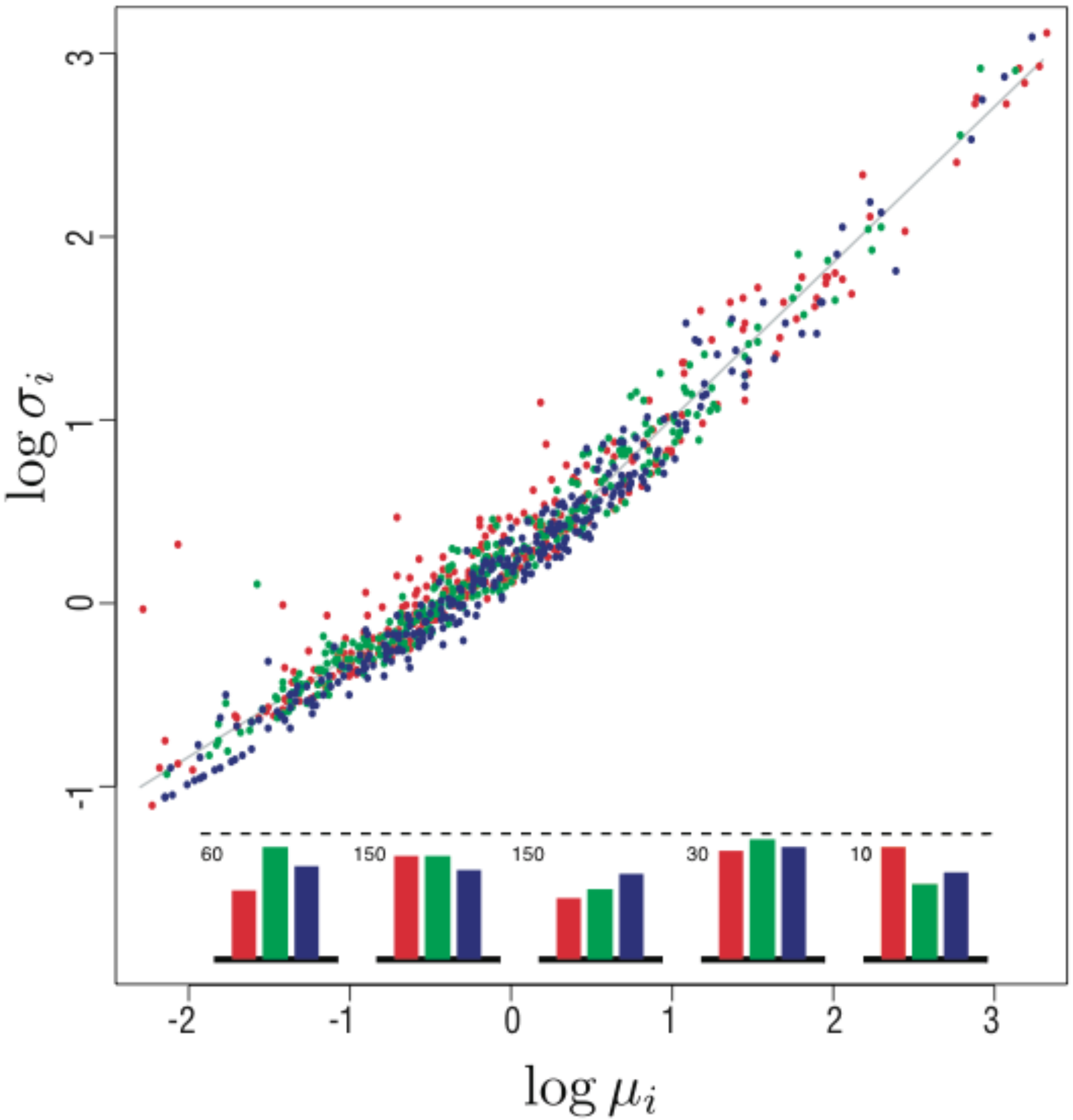


(A) F-statistic smooth and well-behaved. Maximum at $F_{(k)} \approx 1035$ for observation k=1795, corresponding to $\log(\mu_{(1795)}) \approx 0.36$.

(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

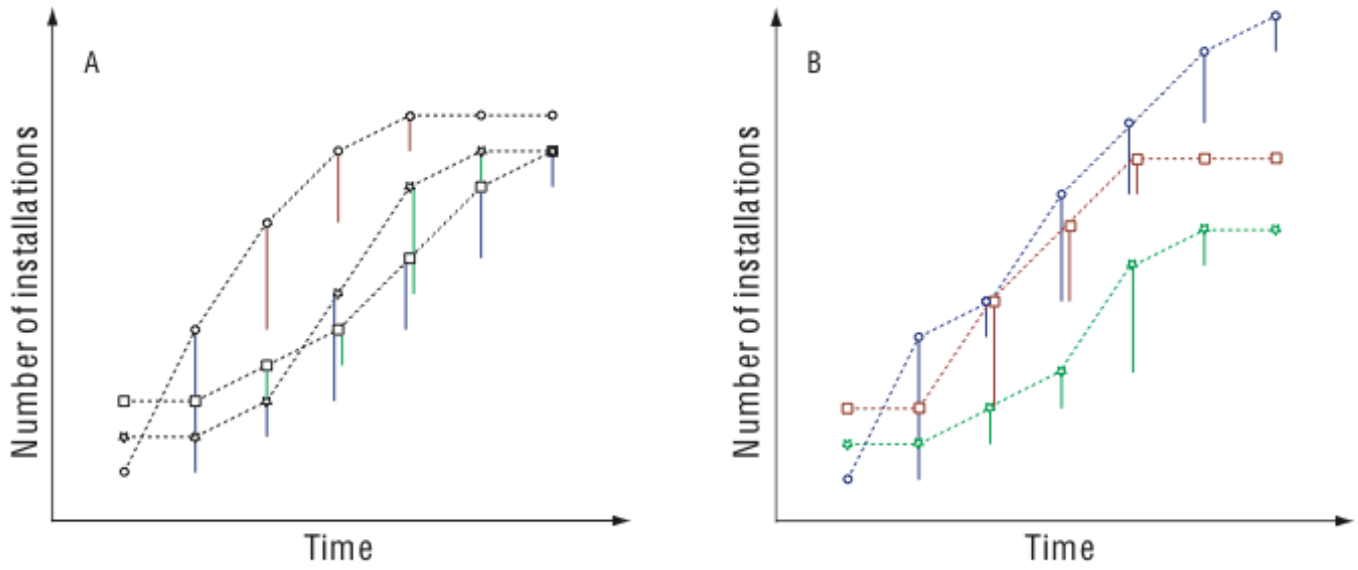
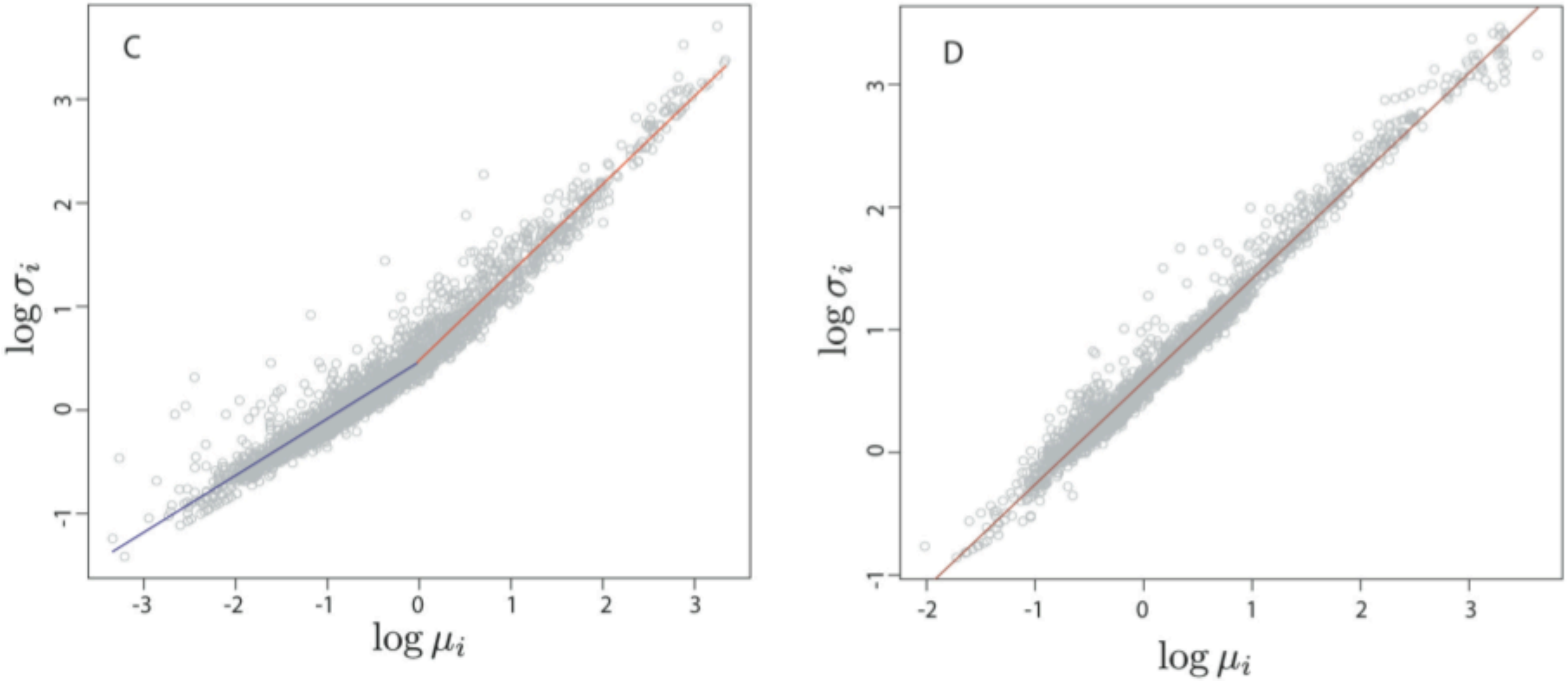


Figure 2: Schematic of the construction of the synthetic time series $\tilde{n}_i(t)$. **(A)** The empirical data consists of $t = 1, \ldots, 7$ observations for three applications. The data points have been connected with dashed black lines to guide the eye. For the most popular application at time $t-1$, 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.

Conclusions to date

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.

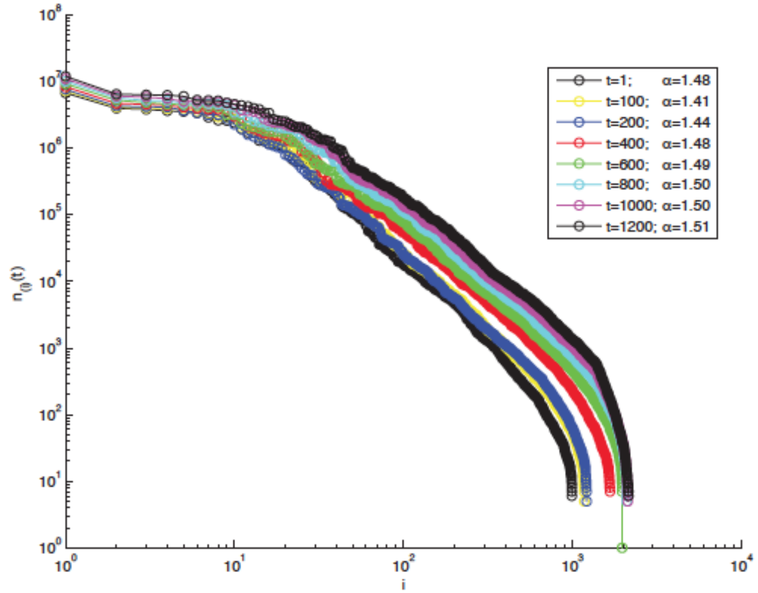
Follow-up work currently in progress:

(i) Developing microscopic models that allow the effects of local and global signals to be distinguished.

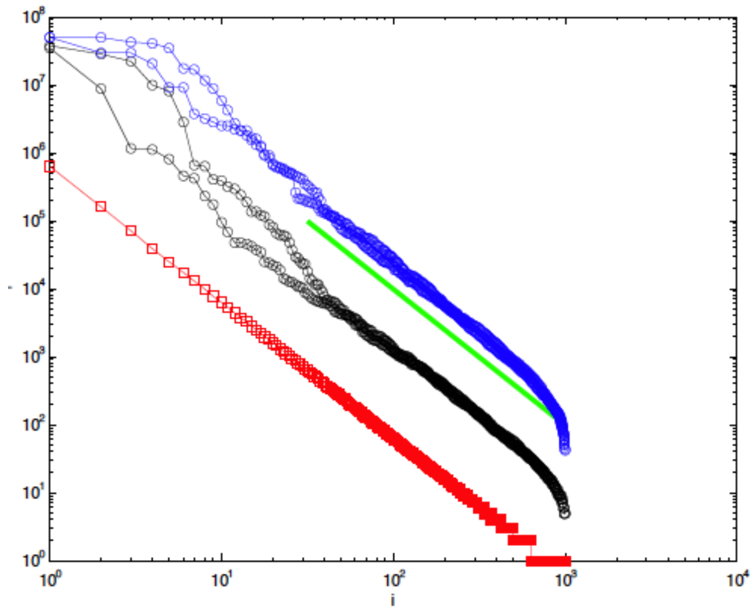
(ii) Applying temporal fluctuation scaling to other online environments.

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

Microscopic models



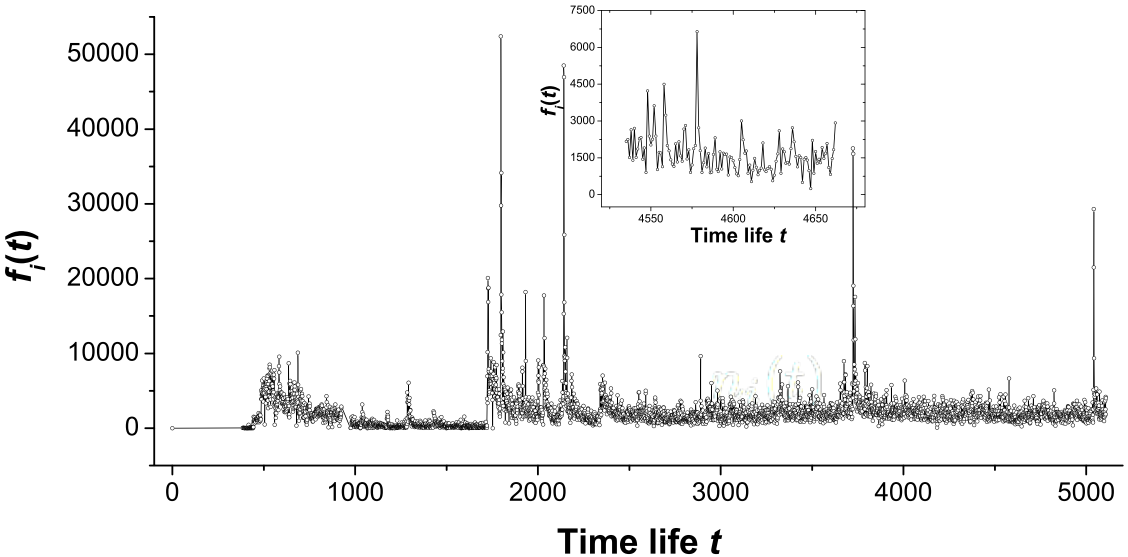
(A) Zipf plot of Facebook application installations in descending order of popularity.

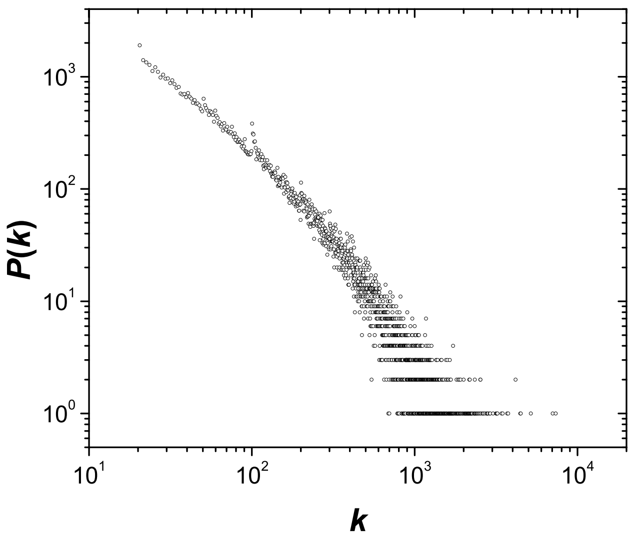


(B) Simulation results.
Red square – model 1 (adoption cascades)
Black circle – model 2 (global influence)

ongoing work with James Gleeson, Mason Porter, Jukka-Pekka Onnela

MovieLens data





http://www.grouplens.org

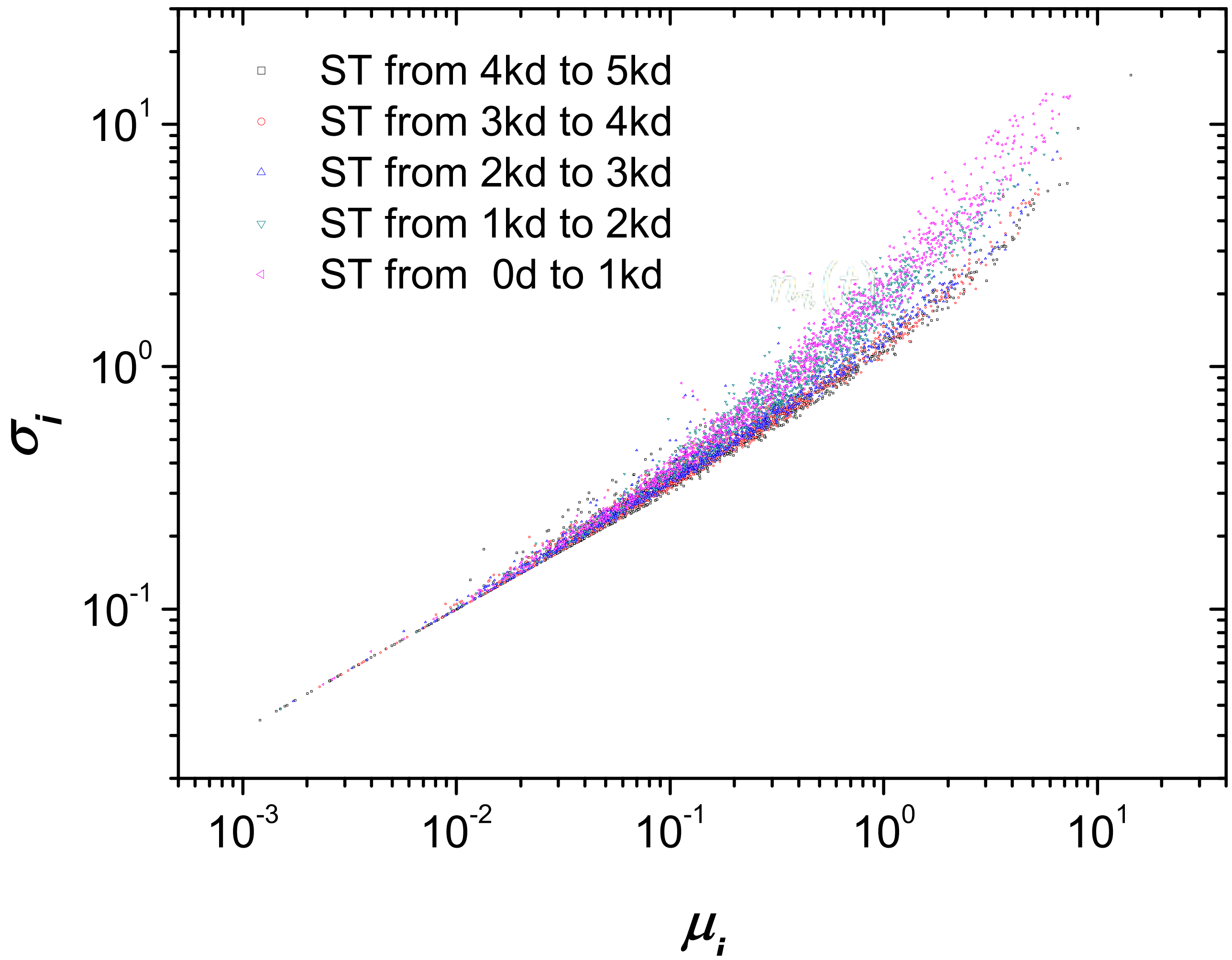
71,567 users

65,133 movies

approx. 20m ratings

9. Jan 1995 – 5 Jan 2009

Fluctuation scaling for movie ratings



ongoing work with Jianguo Liu

Thank you!

Acknowledgements

Co-author: Jukka-Pekka Onnela (Harvard Medical School)

Thanks to:

David Barron (Oxford)
Jerker Denrell (Oxford)
Sean Gourley (Younoodle)
Peter Hedstrom (Oxford)
Nick Jones (Oxford)
Eduardo López (Oxford)
Mason Porter (Oxford)
Jianguo Liu (Oxford)

Zoltan Eisler (Budapest)
James Fowler (UC San Diego)
Gueorgi Kossinets (Google)
Adilson Motter (Northwestern)
Serguei Saavedra (Northwestern)
Matt Salganik (Princeton)
György Szabo (Budapest)
Brian Uzzi (Northwestern)
James Gleeson (Limerick)

EPSRC
Engineering and Physical Sciences
Research Council

SEVENTH FRAMEWORK PROGRAMME

ICTe collective