Twitter's Big Hitters

Des Higham
Department of Mathematics and Statistics
University of Strathclyde

Motivation

Consider large, complex networks with

- a **fixed** set of **nodes**
- **edges** that **come and go over time**

Application areas include

- neuroscience
- communication by email, phone, Twitter
- on-line retail
- on-line social media

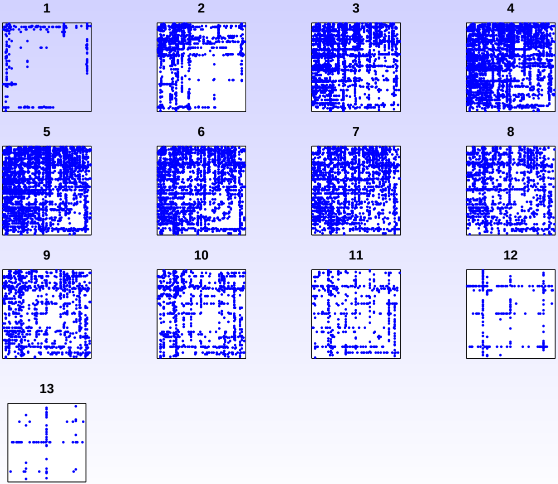
PART 1: **Models**
PART 2: **Algorithms for Computing Centrality**

Example: Voice call contact at MIT

Eagle, Pentland & Lazer, **Proc. Nat. Acad. Sci.**, 2009

106 individuals, 365 days from July 20th, 2004

Summarized into 28 day periods, treated as undirected



Part 1: Models

Challenges

- understand mechanisms
- calibrate parameters and compare models
- forecast future behaviour
- simulate ‘what-if’ scenarios

Grindrod & Higham, **Proc. Royal Society A**, 2010

Discrete time, stochastic models
Given $A^{[k]}$ at time t_k, how do we specify $A^{[k+1]}$?
Think in terms of **birth** and **death** of edges

Triadic Closure Model

Grindrod, Higham & Parsons, **Internet Mathematics**, 2012

Friends of friends become friends

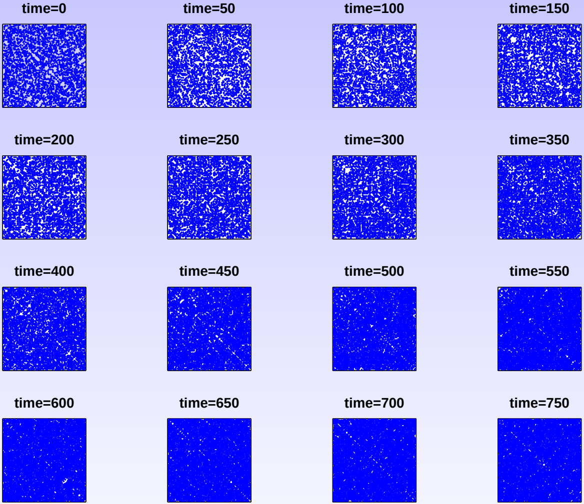
Edge death probability is a constant $\omega \in (0, 1)$
Edge birth probability between nodes i and j given by

$$\delta + \epsilon \left(\left(A^{[k]} \right)^2 \right)_{ij}$$

where $0 < \delta \ll 1$ and $0 < \epsilon(N - 2) < 1 - \delta$

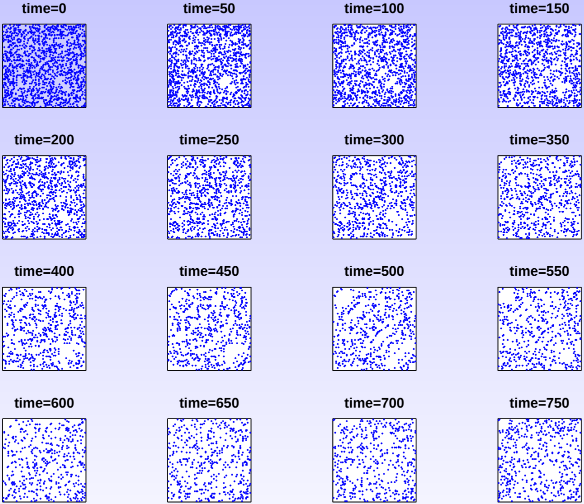
Consider $N = 100$, $\omega = 0.01$, $\epsilon = 5 \times 10^{-4}$, $\delta = 4 \times 10^{-4}$

Triangulation model: start with ER(0.3)



Edge density at $t = 750$ is 0.712

Triangulation Model: start with ER(0.15)



Edge density at $t = 750$ is 0.051

Mean field analysis for $\delta + \epsilon\left(\left(A^{[k]}\right)^2\right)_{ij}$

Ergodicity and **symmetry** $\Rightarrow$ Erdös-Rényi limit: every edge present with probablity $p^\star$

Heuristic **mean field** approach: insert the ansatz "$A^{[k]} = \mathrm{ER}(p_k)$" into the model to obtain

$$p_{k+1} = (1-\omega)p_k + (1-p_k)(\delta + \epsilon(N-2)p_k^2)$$

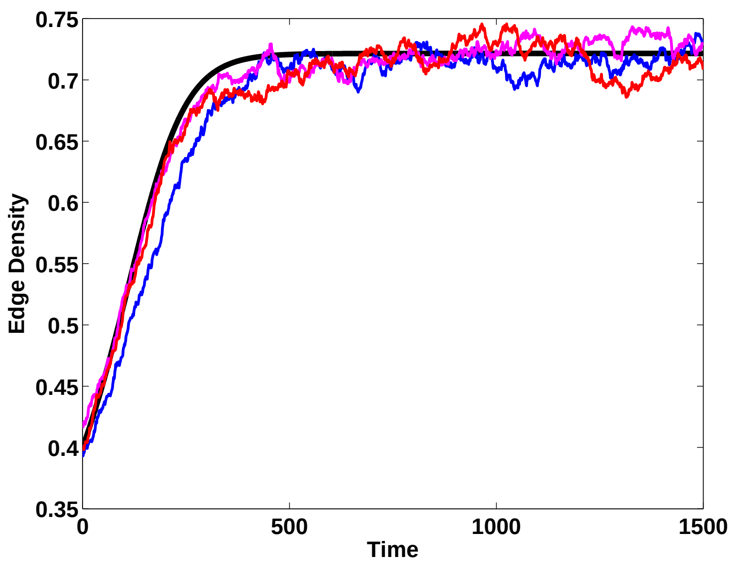
Generically: **three real roots**

Two are stable, **one is unstable**

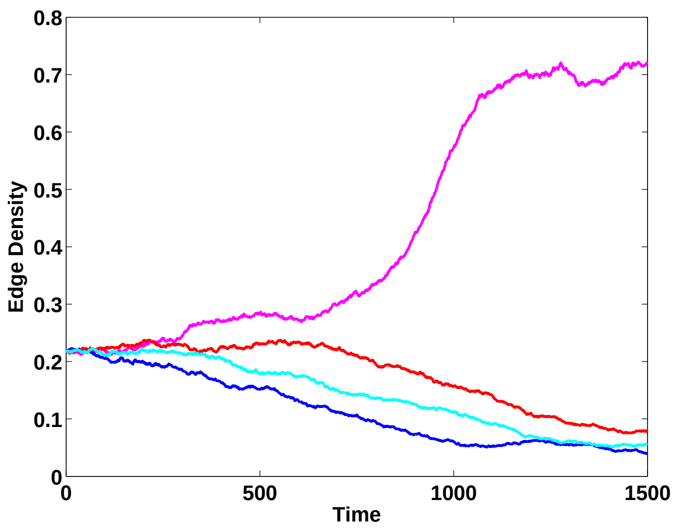
$N = 100$, $\omega = 0.01$, $\epsilon = 5 \times 10^{-4}$, $\delta = 4 \times 10^{-4}$

Stable fixed points 0.049 & 0.721 Unstable 0.229

Mean-field vs. simulation from $\mathrm{ER}(0.4)$



Four simulations from $\mathrm{ER}(0.23)$



Stable fixed points 0.049 & 0.721 Unstable 0.229

Calibration/Inference

Mantzaris & Higham, in *Temporal Networks*, Springer, 2013, edited by P. Holme and J. Saramäki

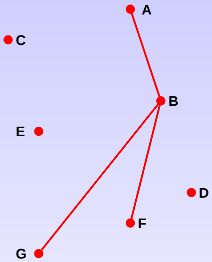
Given model parameters, we can compute the probability of observing the data: **likelihood**

Tests on synthetic data show that we can correctly infer the triadic closure effect

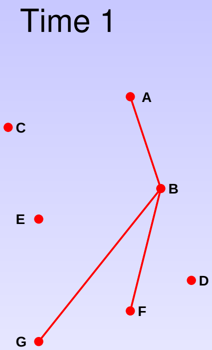
Wealink data from Hu and Wang, **Phys. Lett. A**, 2009.
26 Million time stamps, over 841 days with 0.25 Million nodes (no edge death):
we found statistical support for triadic closure

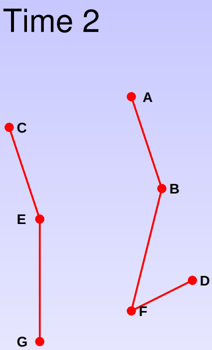
Part 2: Algorithms for Node Centrality

Time 1 Time 2 Time 3



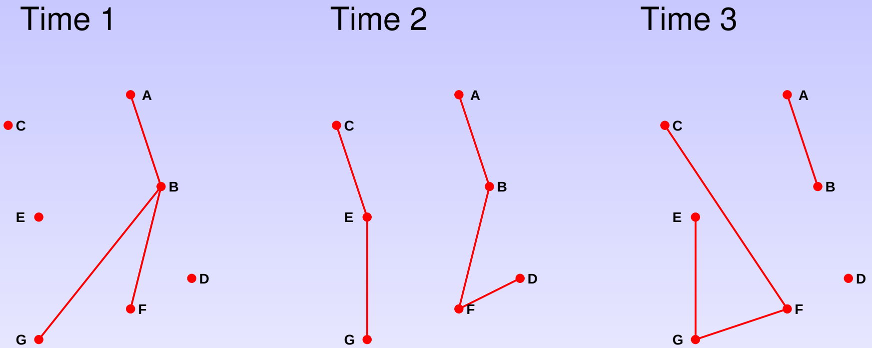
Part 2: Algorithms for Node Centrality



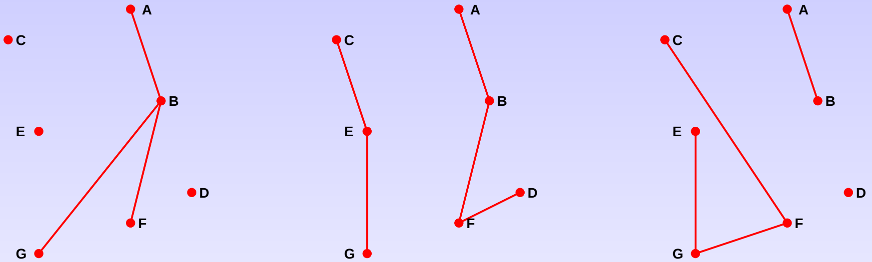


Time 3

Part 2: Algorithms for Node Centrality



Time 1 Time 2 Time 3



Note the lack of symmetry caused by time's arrow

Dynamic Walks

Time points $t_0 < t_1 < t_2 < \cdots < t_M$

Adjacency matrices $A^{[0]}, A^{[1]}, A^{[2]}, \ldots, A^{[M]}$

Dynamic walk of length w from node i_1 to node i_{w+1}:
sequence of times $t_{r_1} \leq t_{r_2} \leq \cdots \leq t_{r_w}$ and a
sequence of edges $i_1 \leftrightarrow i_2, i_2 \leftrightarrow i_3, \ldots, i_w \leftrightarrow i_{w+1}$,
such that $i_m \leftrightarrow i_{m+1}$ exists at time t_{r_m}

(Several variations are possible)

Use this to define **centrality** of a node, following **Katz**[1]

[1] *A new status index derived from sociometric analysis*, Leo Katz, **Psychometrika**, 1953

New Algorithm

Grindrod, Higham, Parsons & Estrada, **Phys. Rev. E**, 2011

Key observation: the matrix product

$$A^{[r_1]}A^{[r_2]}\dots A^{[r_w]}$$

has i, j element that counts the number of dynamic walks of length w from node i to node j, where the mth step takes place at time t_{r_m}

Keep track of all such walks and discount by α^w

E.g. $\alpha^2 A^{[0]}A^{[1]}$, $\alpha^4 A^{[0]}A^{[2]}A^{[3]}A^{[7]}$, $\alpha^3 A^{[3]}A^{[3]}A^{[9]}$

This is achieved by

$$\mathcal{Q} := \left(I-\alpha A^{[0]}\right)^{-1}\left(I-\alpha A^{[1]}\right)^{-1}\cdots\left(I-\alpha A^{[M]}\right)^{-1}$$

Then $\mathcal{Q}$ is our overall summary of how well information can be passed from node i to node j

Dynamic Centrality

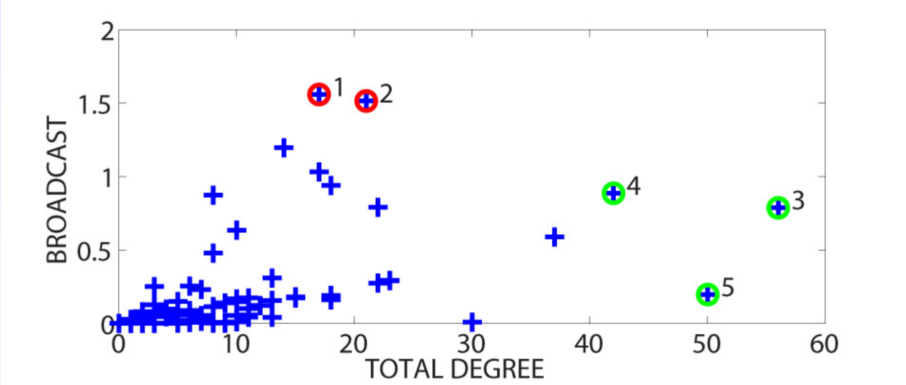
We will call the row and column sums

$$\sum_{k=1}^{N} \mathcal{Q}_{nk} \qquad \& \qquad \sum_{k=1}^{N} \mathcal{Q}_{kn}$$

the **broadcast** and **receive** communicabilities

- generalizes **Katz centrality** in social networks
- involves **sparse linear solves**
- captures the **asymmetry** through non-commutativity of matrix multiplication

Enron email: 151 nodes over two weeks



Explanatory model for **dynamic communicators** in:

Mantzaris & Higham, **Eur. J. Appl. Math.**, 2012

fMRI in Neuroscience: functional connectivity

Mantzaris, Bassett, Wymbs, Estrada, Porter, Mucha, Grafton, Higham, **J. Complex Networks**, 2013

Data: Bassett, Wymbs, Porter, Mucha, Carlson, Grafton, **PNAS**, 2011

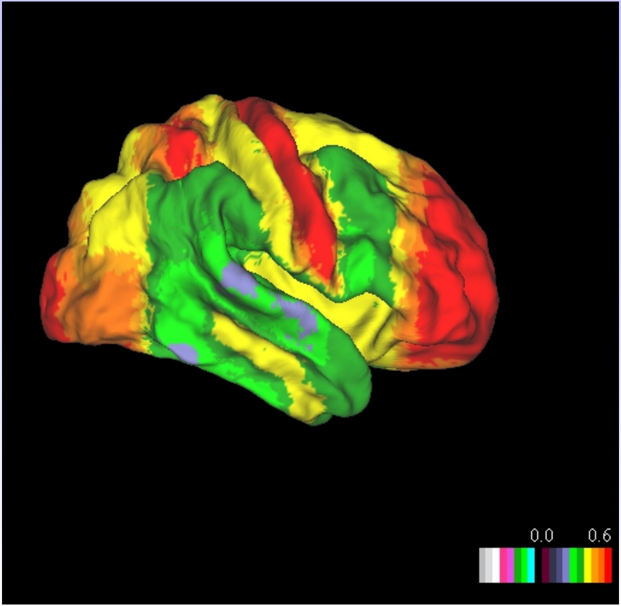
Each experiment: $112 \times 112 \times 25$ tensor
region region time

60 experiments: 20 subjects repeat a task three times

Unsupervised k-means custering of the full data set shows **significant evidence for "learning"**:
subjects typically move from cluster 1 to cluster 2

Summarizing each experiment in terms of the 112 broadcast (or receive) centrality measures, we recover the **same level of significance**.
And we can now **interpret the data more easily**......

Change in broadcast centrality: Right Medial



Twitter's Big Hitters

Laflin, Mantzaris, Grindrod, Ainley, Otley, Higham,
Proc. of Social Informatics, 2012

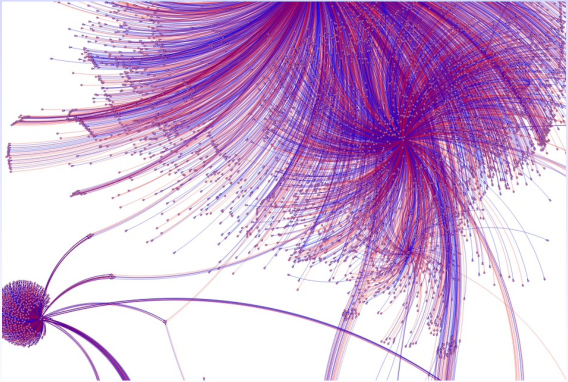
Listen to **tweets** containing the phrases `city break`, `cheap holiday`, `travel`, `insurance`, `cheap flight` plus two brand names

From 17 June 2012 at 14:41 to 18 June at 12:41

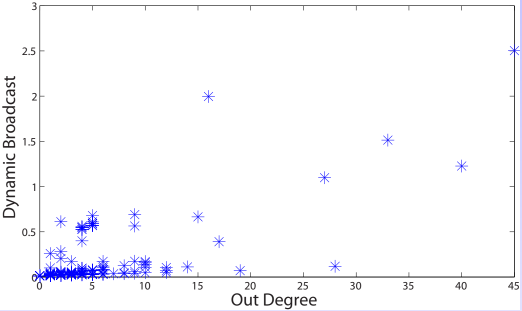
0.5 Million Tweeters/Followers

Active Node Subnetwork Sequence

- Record all relevant edges: tweeter $\mapsto$ followers
- Remove all nodes with zero out degree
- Binarize over time windows of length Δt



Dynamic Broadcast Centralities

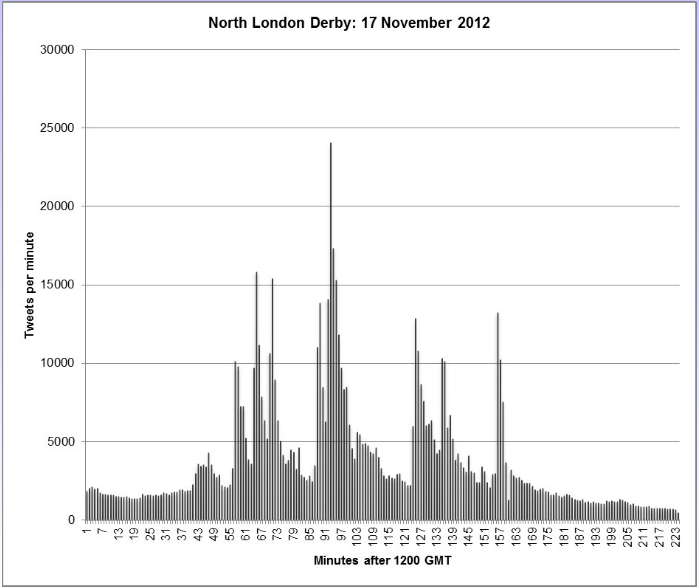


Twitter account with **fourth highest bandwidth** (out degree) is a **very poor dynamic broadcaster**
Closer inspection ⇒ an automated process.

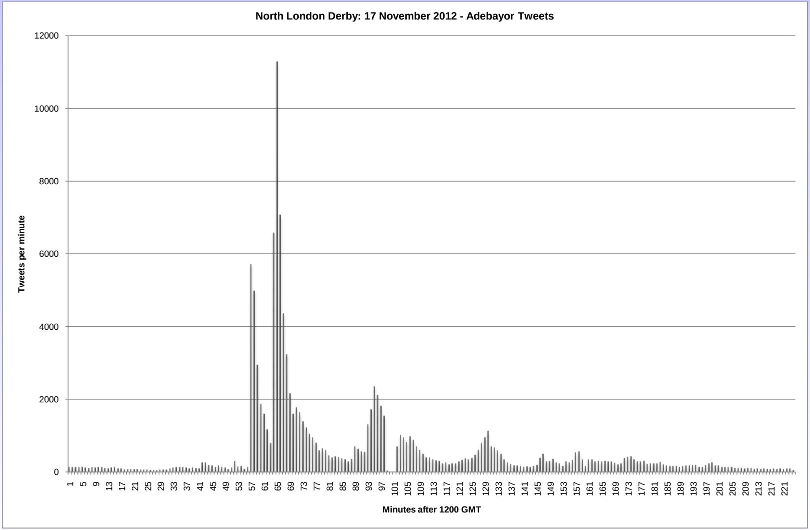
Five **social media experts** were given the Twitter data and asked to rank the accounts according to importance

We found that **dynamic centrality measures are hard to distinguish from human experts**

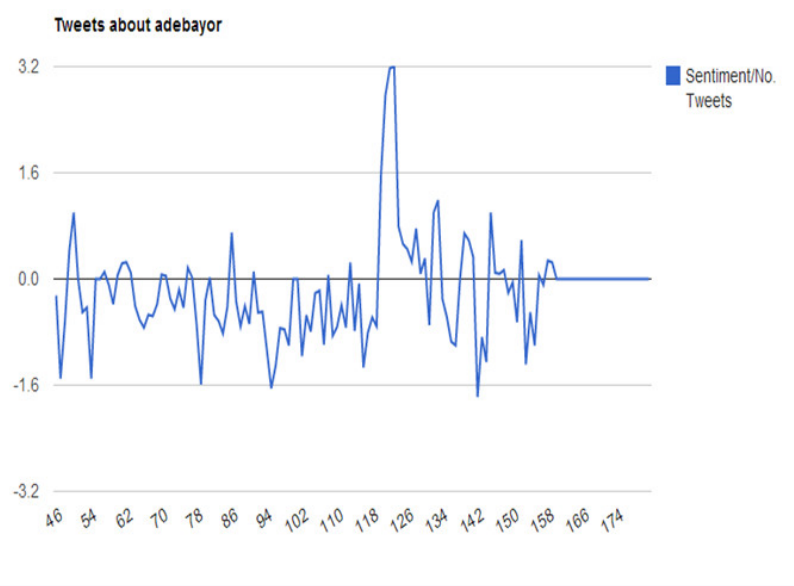
Arsenal 5 - 2 Spurs, November 2012



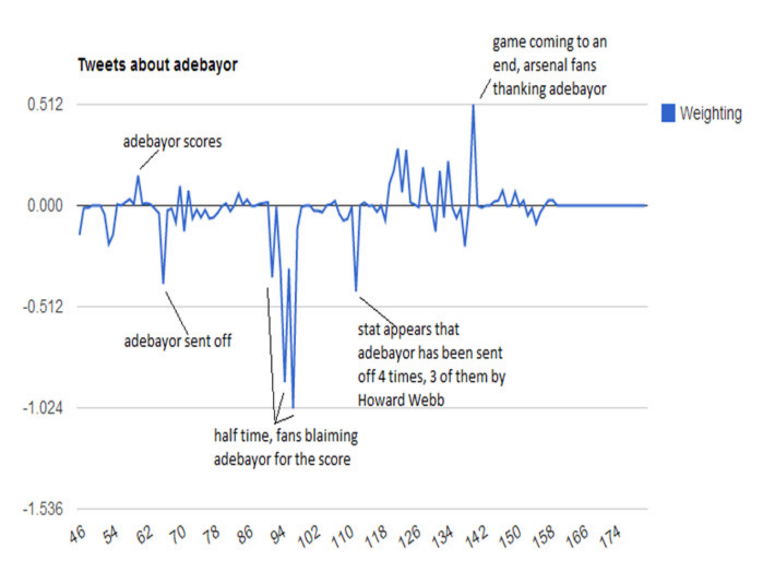
Adebayor: volume of tweets



Adebayor: sentiment across time



Adebayor: sentiment weighted by influence



Downweighting over time

Grindrod & Higham, **SIAM Review** (Research Spotlights) 2013

Motivation: News goes stale, messages become irrelevant, viruses mutate, ... ***old information is less important***

The algorithm can be generalized naturally to

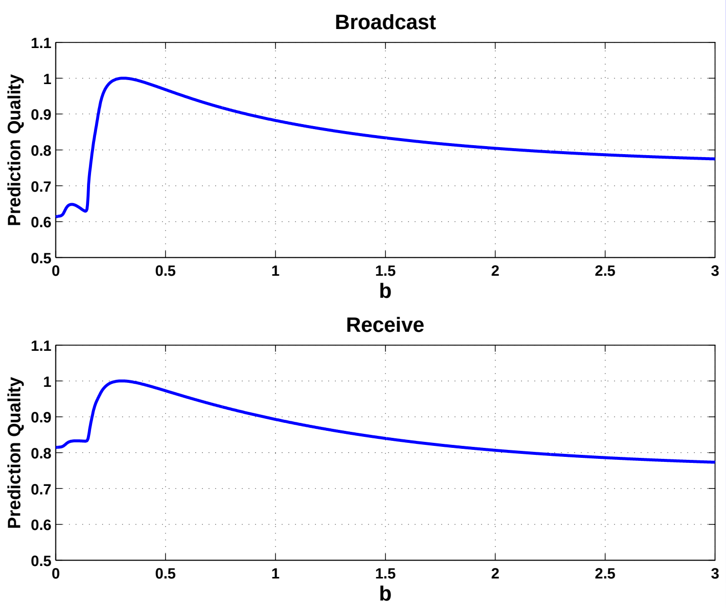
$$\mathcal{S}^{[k]} = \left(I + e^{-b\Delta t_k}\mathcal{S}^{[k-1]}\right)\left(I - \alpha\, A^{[k]}\right)^{-1} - I$$

Here, $\left(\mathcal{S}^{[k]}\right)_{ij}$ counts the number of dynamic walks from i to j up to time t_k, scaled by

- a factor α^w for **dynamic walks of length w**
- a factor e^{-bt} for **walks that begin t time units ago**

We have a new parameter, b:
$b = 0$ is the previous algorithm
$b = \infty$ is Katz on the current network

Enron daily emails: Centrality prediction



Continuous Time?

Discretizing in time and binarizing is convenient, but

- Δt too large can **overlook** or **smear** events
- Δt too small may give a **false impression** of accuracy

So create $A(t)$ over continuous time

$\Rightarrow$ ODE for the evolution of pairwise communicability:

$$U'(t) = -b(U(t) - I) - U(t) \log \left(I - aA(t) \right)$$

What's New?

- Edge birth model based on **triadic closure**
- Mean field analysis that accurately summarizes the macro scale behavoiur and predicts **bistability**
- Concept of **dynamic walks**
- **Resolvent-based algorithms** generalizing **Katz**
- **Continuous-time** analogues

What's Next?

- Large scale **calibration/inference/model comparison**
- Dynamic network **monitoring/prediction/disruption**
- Algorithms for **dynamic clustering/classification**, e.g., fake account detection

Thanks to
EPSRC/RCUK Digital Economy programme,
Leverhulme Trust,
EPSRC/Strathclyde Impact Acceleration Account,
Bloom Agency.