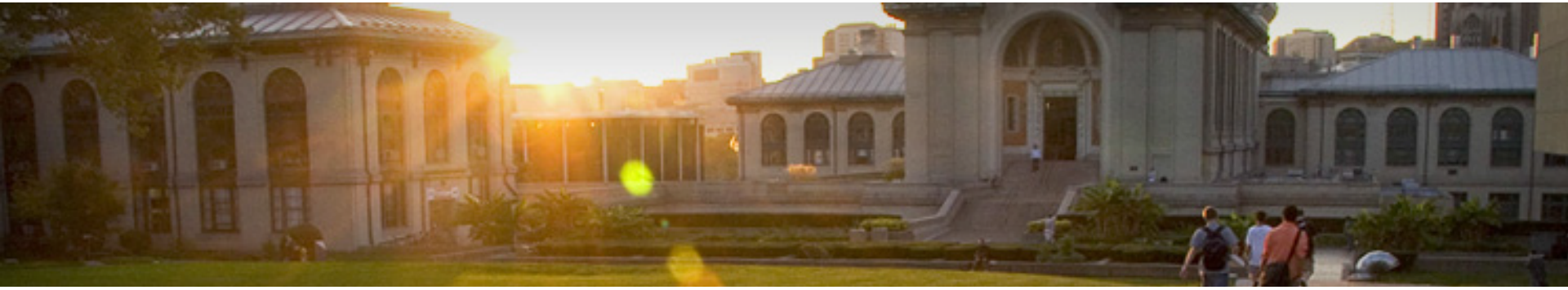
Carnegie Mellon University
School of Computer Science



Platform effects in social media networks

Momin M. Malik[1] & Jürgen Pfeffer[2]
[1]Institute for Software Research, Carnegie Mellon University
[2]Bavarian School of Public Policy, Technical University of Munich

Second International Conference on Computational Social Science (IC^{2}S^{2}2016)
Kellogg School of Management at Northwestern University, Evanston, Illinois
Session: Social Networks 1
June 24, 2016

These slides available at:
http://mominmalik.com/iccss2016.pdf

Full paper in *ICWSM-16*:
http://mominmalik.com/icwsm2016.pdf

Overview

Three key recommendations for computational social science:

1. **Social media network data are not neutral measurements** (just like every scientific instrument)
2. Data artifacts can be opportunities, not annoyances
3. There are relevant observational inference methods

Measurement, instruments and science

"Disciplines are revolutionized by the development of novel tools: the telescope for astronomers, **the microscope for biologists,** the particle accelerator for physicists, and brain imaging for cognitive psychologists. **Social media provide a high-powered lens into the details of human behavior and social interaction** that may prove to be equally transformative."

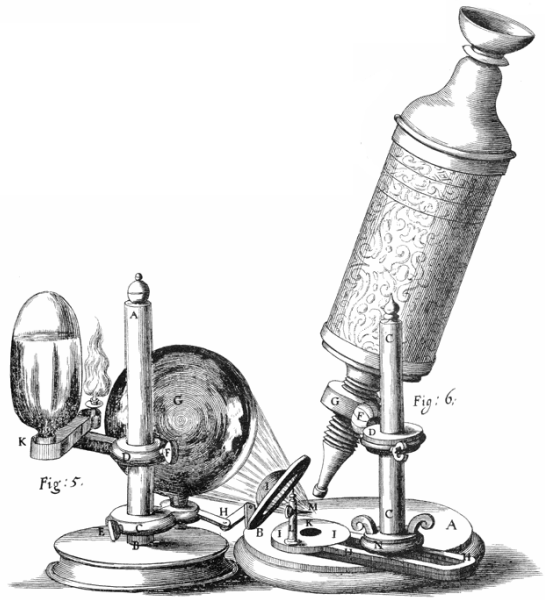
Golder, S., & Macy, M. (2012). Social science with social media. *ASA footnotes*, *40*(1), 7.

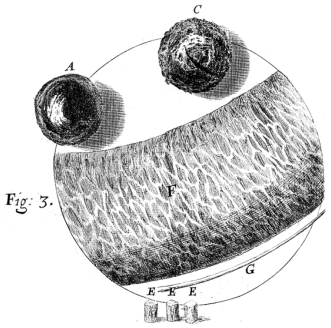
A "cell theory" for social networks?

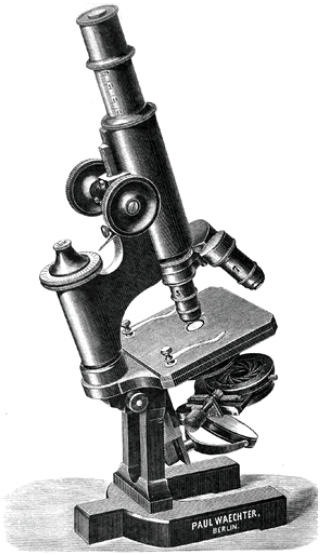
"For example, **what does existing sociological network theory,** built mostly on a foundation of one-time "snapshot" data, typically with only dozens of people, **tell us about massively longitudinal data sets of millions of people,** including location, financial transactions, and communications? These **vast, emerging data sets on how people interact surely offer qualitatively new perspectives** on collective human behavior, but our current paradigms may not be receptive."

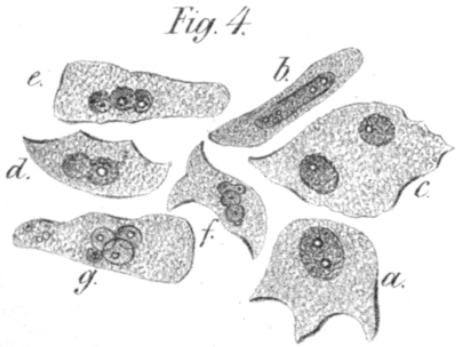
Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., Alstyne, M. A. (2009) Computational social science. *Science*, *323*(5915), 721–723.

But: cells *described* in 1665; *cell theory* in 1830s!









Hooke, R. (1665). *Micrographia: or some physiological descriptions of minute bodies made by magnifying glasses. With observations and inquiries thereupon*. London: J. Martyn and J. Allestry.
Virchow, R. (1847). Ueber die Standpunkte in der wissenschaftlichen Medicin. *Archiv für Pathologische Anatomie und Physiologie. 1*(1), 3-19.

Why the 165-year "delay"?

- Improving the *illumination*, more than improving resolution, was the key
- Rudolf Virchow's work establishing cell theory in 1830s was also aided by his use of *staining*
- I.e., treating the phenomena to be amenable to the instrument is as important as the instrument

Szekely, F. (2011). Unreliable observers, flawed instruments, 'disciplined viewings': Handling specimens in early modern microscopy. *Parergon, 28*(1).
Laboratory for Optical and Computational Instrumentation. (2015). History of the light microscope. Microscopy Museum, University of Wisconsin-Madison. http://loci.wisc.edu/outreach/history-light-microscope
Epstein, B. (2015). *The ant trap: Rebuilding the foundations of the social sciences*. Oxford University Press.

Understand the biases in the data

Networks in Twitter, Instagram, Facebook, Reddit, 4chan...

- Differences in demographics, adoption patterns
- Heterogeneous users (e.g., corporate users, celebrity fans, activists)
- Norms of use, platform-specific culture and behavior
- Access constraints, platform-side filtering
- International differences
- Changes over time

These make results not generalize!!

Gayo-Avello, D. (2011). Don't turn social media into another 'Literary Digest' Poll. *Communications of the ACM*, *54*(10), 121-128.
Tufecki, Z. (2014). Big questions for social media big data: Representativeness, validity, and other methodological pitfalls. *ICWSM-14*, 505-514.
Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, 346(6213), 1063-1064.
Harigttai, E. (2015). Is bigger always better? Potential biases of big data derived from social network sites. *Annals of the American Academy of Political and Social Science*, *659*, 63-76.

Remaining blind spot: Companies themselves

- Social media platforms are not neutral utilities nor research environments
- They are corporations, concerned about markets and their business model
- They are constantly engineering and re-engineering platforms to *encourage desirable behavior* from their users, e.g.,
  - Spending more time on the platform
  - Producing certain kinds of data
- Simple example: 140 character limit *forces* human output into standardized units that are far easier to store, process, and link

van Dijjck, J. (2013). *The culture of connectivity: A critical history of social media*. Oxford University Press.
Gehl, R. W. (2014). *Reverse engineering social media: Software, culture, and political economy in new media capitalism*. Temple University Press.

We call the success of such engineering attempts *platform effects*.

Key question: *How much of observed behavior is due to platform effects?*

Corollary: Are we studying user behavior? Or just the successful algorithmic management of users by industry engineers?

Data artifacts as natural experiments

- A/B tests from inside companies may have all the answers, but not accessible from outside
- Instead, find *natural experiments*.
- Discontinuities that result from platform changes: typically seen as "data artifacts," annoyances incidental to what we want to study
- But we can exploit them them as potential natural experiments for platform effects!

Example: Netflix movie ratings

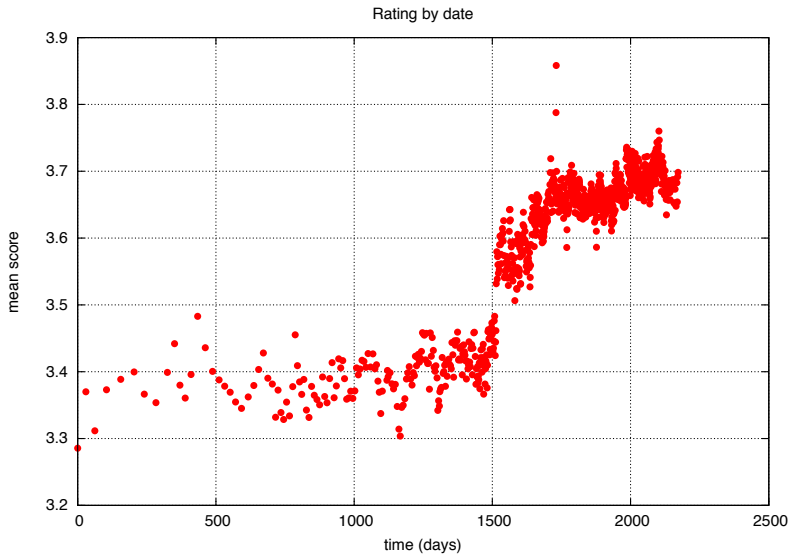
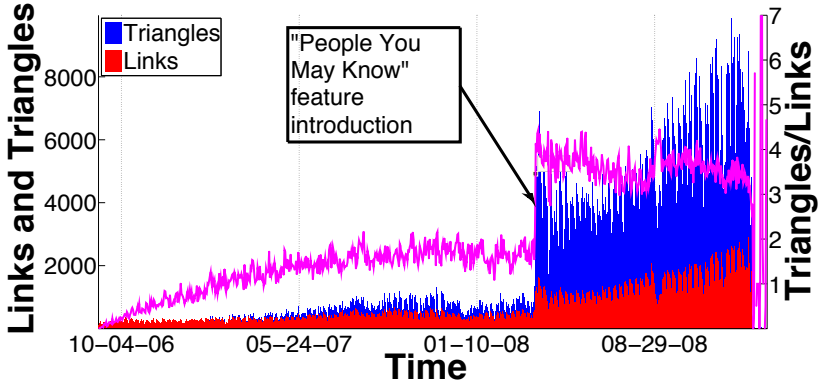


Fig. 1 from Koren (2009): Temporal effects emerging within the Netflix movie rating dataset. The average movie rating made a sudden jump in early 2004 (1500 days since the first rating in the dataset). Each point averages 100,000 rating instances.

"This hints that beyond a constant improvement in matching people to movies they like, something else happened in early 2004 causing an overall shift in rating scale..."

Koren, Y. (2009). Collaborative filtering with temporal dynamics. *KDD '09*, 447–456.

Example: Facebook New Orleans



(b) Facebook New Orleans

Fig. 2 from Zignani et al. (2014): Number of new links (red) and triangles (blue) formed during the growth of Facebook New Orleans, sampled each day. The magenta line represents the ratio between the triangle and the links created in a day (y-scale on the right).

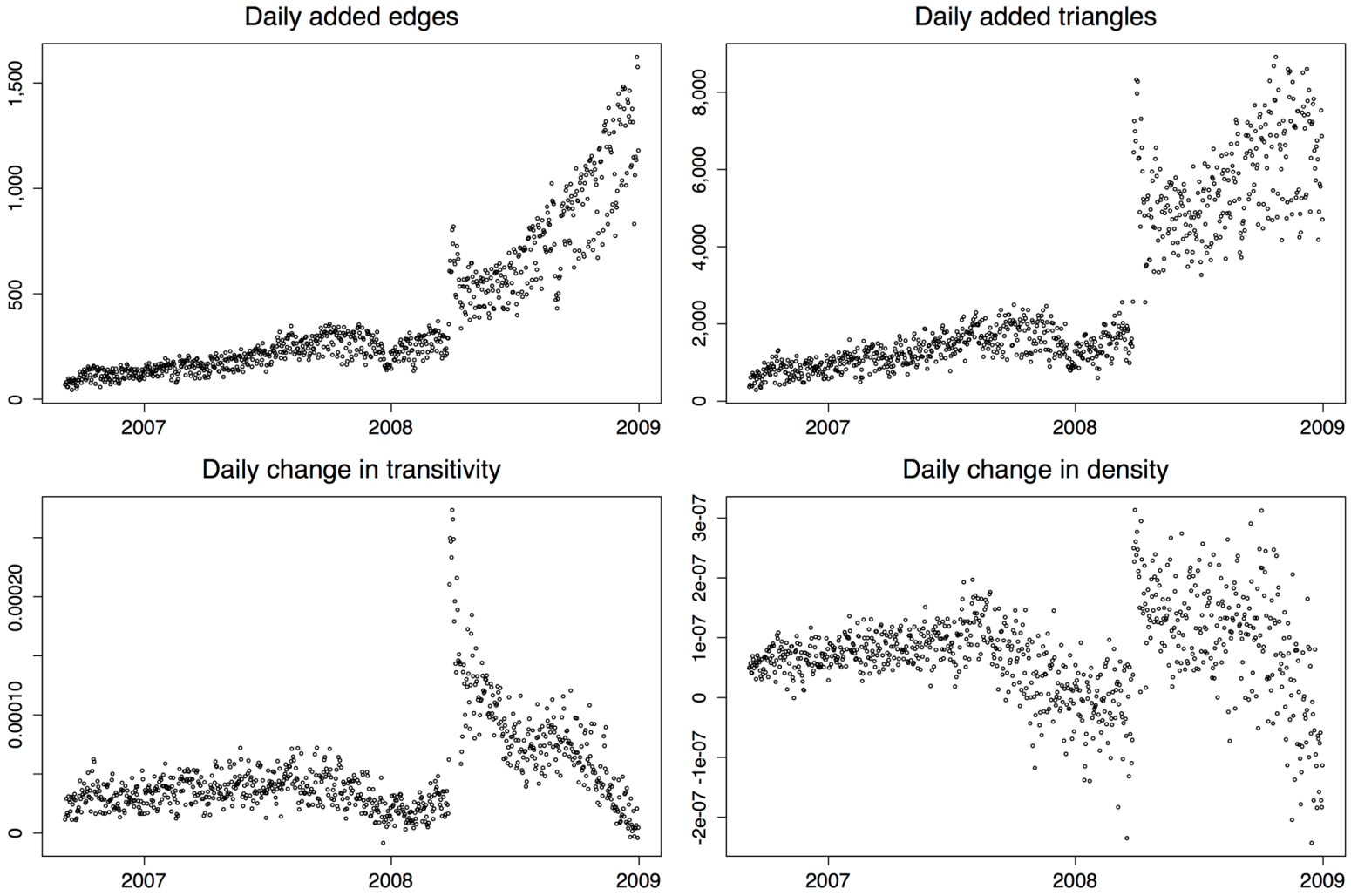
Data set crawled by Viswanath et al. (2009)

- Publically available profiles in FB New Orleans, ~52% of all profiles there (multiple issues in boundary specifications, other issues in data)
- About 800,000 unique edges

Zignani et al. (2014) notice a strange spike...

Viswanath, B., Mislove, A., Cha, M., & Gummadi, K. P. (2009). On the evolution of user interaction in Facebook. *WOSN '09*, 37-42.
Zignani, M., Gaito, S., Rossi, G. P., Zhao, X., Zheng, H., & Zhao, B. Y. (2014). Link and triadic closure delay: Temporal metrics for social dynamics. *ICWSM-14*, 564-573.

Facebook New Orleans time series



Causal estimation

- $Y_i|set(T=1)$ is value of Y_i if i given treatment T
  $Y_i|set(T=0)$ is value of Y_i if i not given treatment T
- For a given i, can never observe both
- Instead, use expectations. Define the *average treatment effect* α as

$$\alpha := E(Y_i|set(T=1)) - E(Y_i|set(T=0))$$

- If $\{Y_i|set(T=1), Y_i|set(T=0)\}$ is independent of of T (what randomization does), then

$$E(Y_i|set(T)) = E(Y_i|T)$$

Regression discontinuity

Regression Discontinuity (RD) Design is the use of a treatment that is effective strictly above some cutoff value c of a covariate X_i, $T = \mathbf{1}(X_i > c)$.

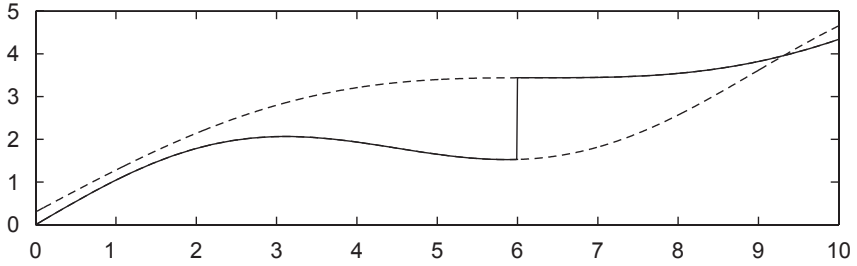


Fig. 2 from Imbens and Lemieux (2008): Potential and observed outcome regression functions.

Point estimate of the effect of treatment on the treated is the *local average treatment effect,*

$$\alpha = E(Y_i|\mathbf{1}(X_i > c) = 1) - E(Y_i|\mathbf{1}(X_i > c) = 0)$$
$$= \lim_{x \downarrow c} E(Y_i|X_i = x) - \lim_{x \uparrow c} E(Y_i|X_i = x)$$

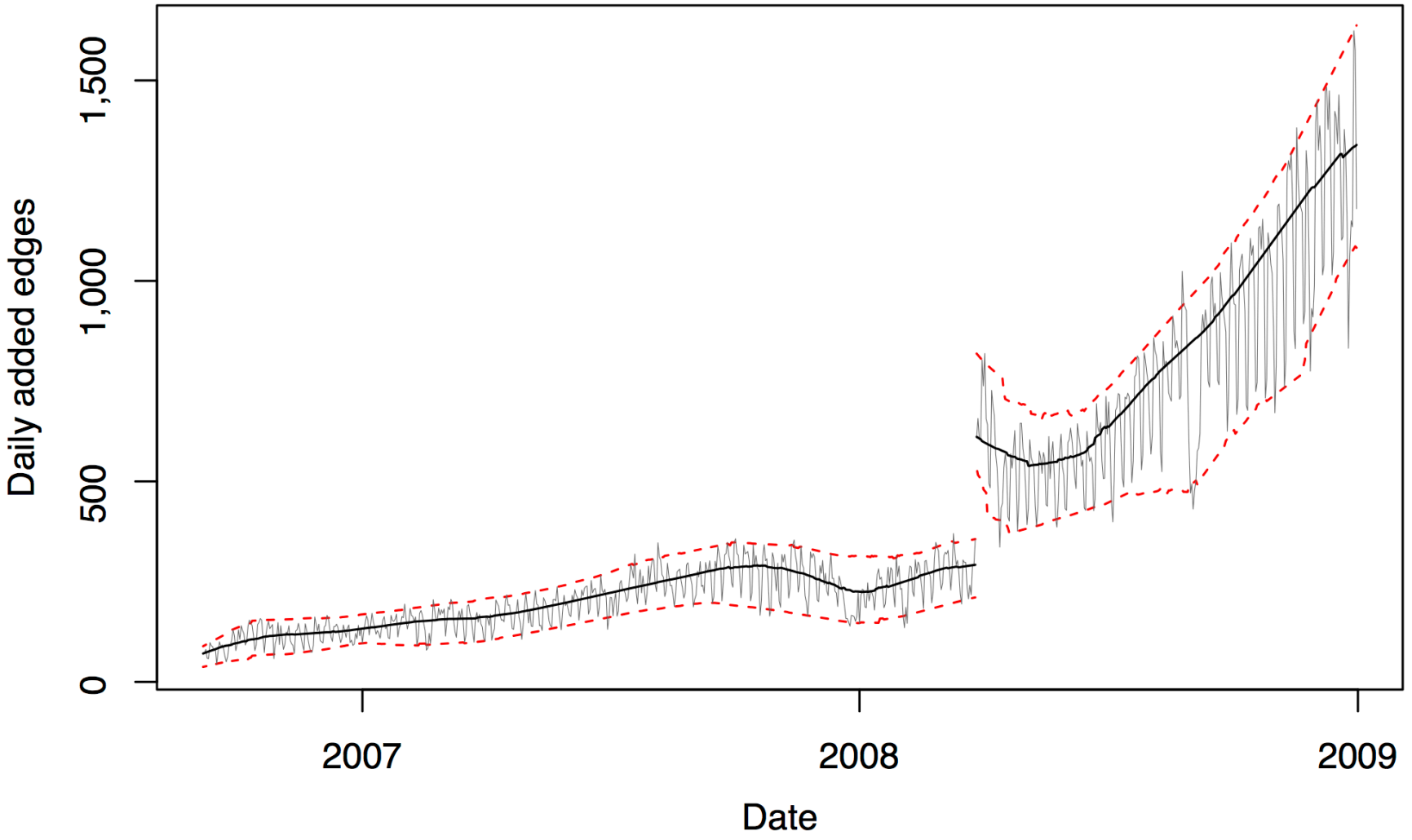
Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*(2), 615–635.
Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica, 62*(2), 467– 475.

Time series specification testing

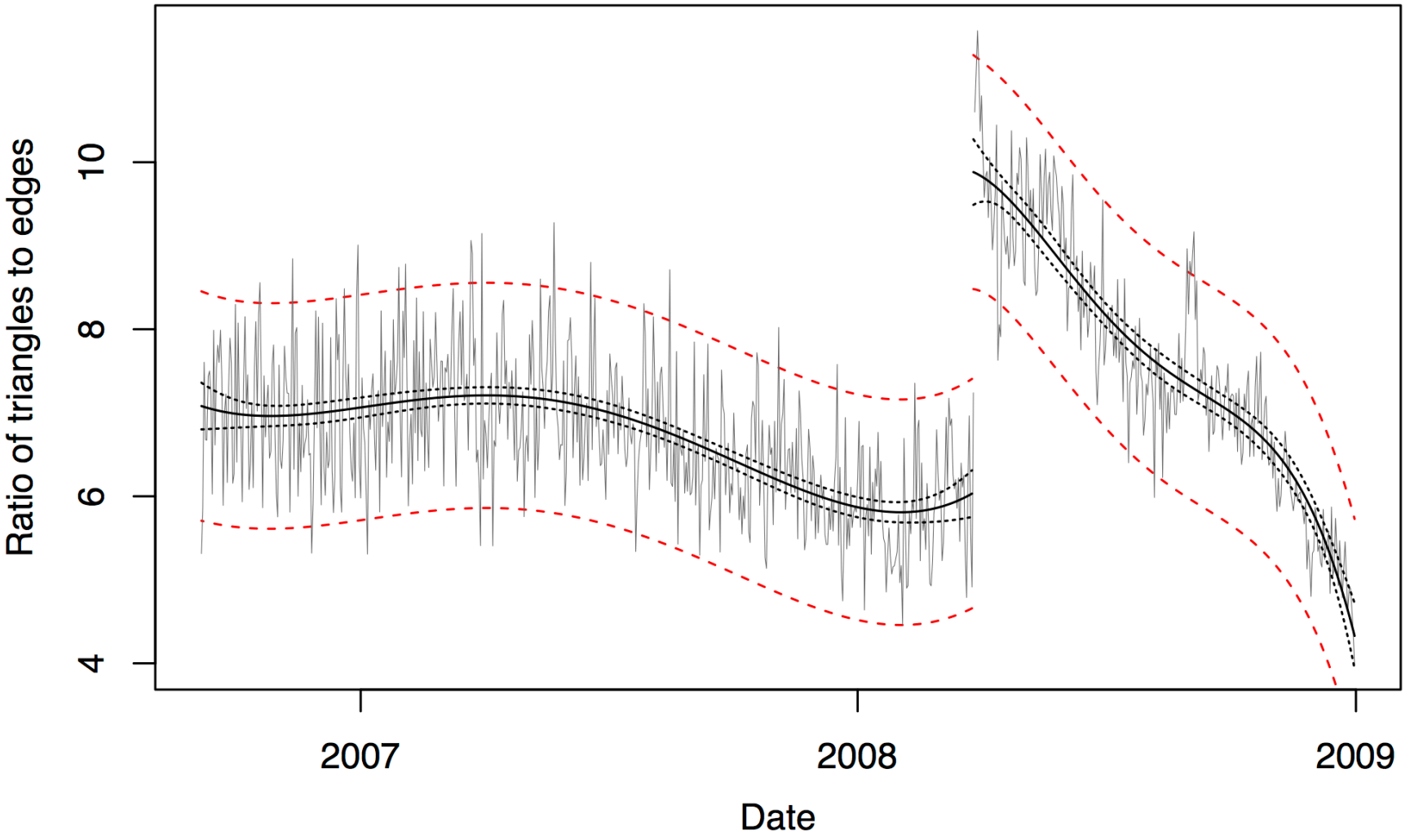
- Want to test, is this really a discontinuity? See if, for other choices of a cutoff, the confidence intervals to the left and right overlap
- Problem: RD not made for time series. Not accounting for temporal autocorrelation makes confidence intervals too narrow
- Alternatives?
  - Event Analysis, Interrupted Time Series: not as formal as RD
  - ARIMA models: Differencing destroys discontinuity (are ways to do, but we preferred RD as a full causal framework)
  - Gaussian Process regression: CIs still too narrow
- Solution: tolerance intervals (empirical coverage), which we fit with quantile regression. Captures irreducible variance of time series, gives wide enough intervals

Two key findings

Facebook: ~+300 edges per day (+200%)!



Facebook: ~+3.8 daily triangles per edge!



Limitations

- Time series of network statistics are not specifically network models; and, first differencing not enough to make the time series of observed network statistics stationary
- Major questions about data quality (not critical)
- Are raw ties the right thing to measure? (E.g., measure *interaction* instead?)
- Next, look at some metrics related to celebrated computational social science results: average path length vs. clustering coefficient, degree distribution, diameter…

Conclusions

Study platform effects! They change what we will think is happening.

"Data artifacts" are not curiosities or annoyances, they can be deeply theoretically revealing about platform effects

As external researchers, apply observational inference techniques to data artifacts for researching platform effects

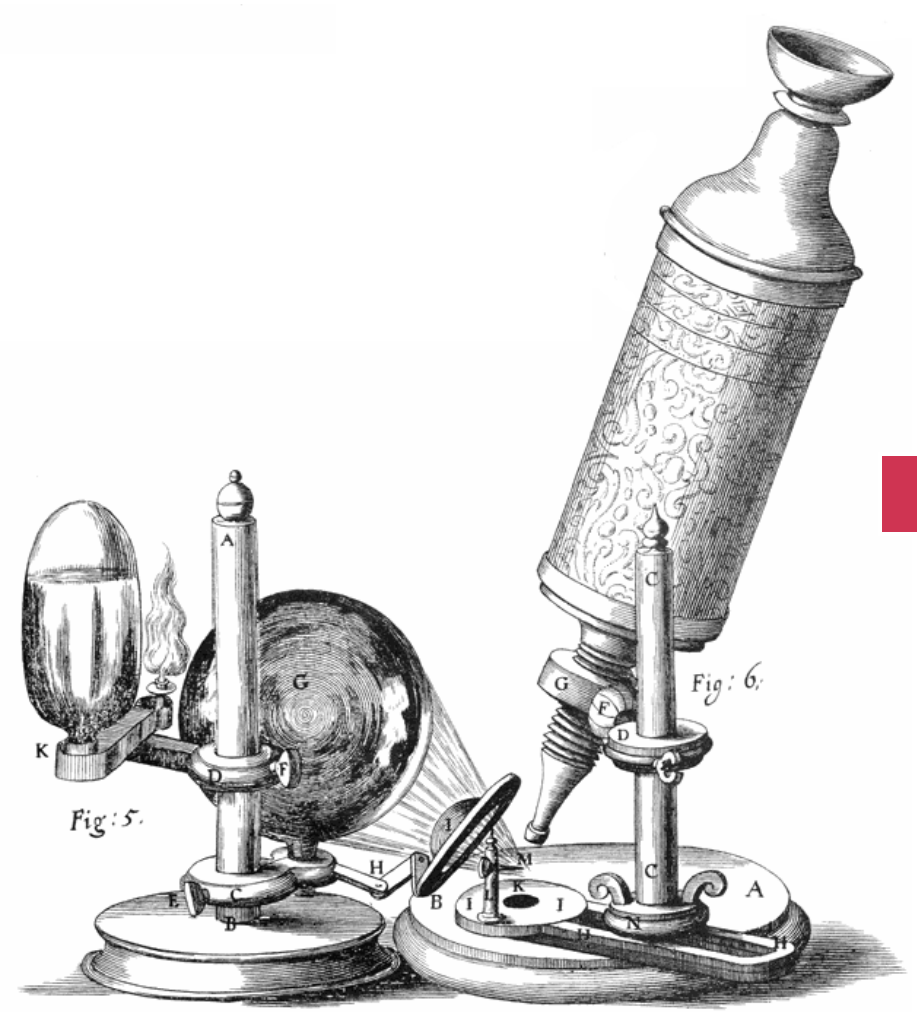
And...

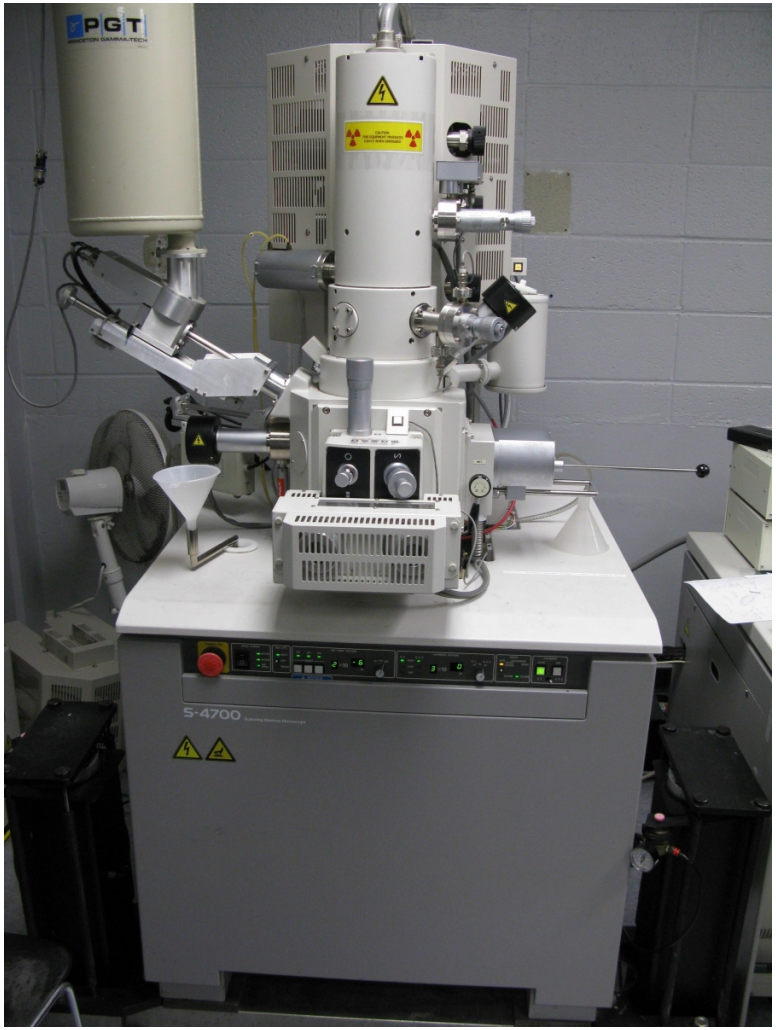
Don't mistake what the instrument measures for the underlying phenomenon!

In order to know what/how to measure, we already have to have a pretty good idea of what we are looking for

Much of science involves improving the tools and how we use them, as we understand what we are trying to study

Not yet a high-power lens, but can be





Thank you! Questions?

momin@cmu.edu
These slides: http://mominmalik.com/iccss2016.pdf
Full paper: http://mominmalik.com/icwsm2016.pdf