Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

Identifying Platform Effects in Social Media Data

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

Abstract

Even when external researchers have access to social media data, they are not privy to decisions that went into platform design—including the measurement and testing that goes into deploying new platform features, such as recommender systems, seeking to shape user behavior towards desirable ends. Finding ways to identify platform effects is thus important both for generalizing findings, as well as understanding the nature of platform usage. One approach is to find temporal data covering the introduction of a new feature; observing differences in behavior before and after allow us to estimate the effect of the change. We investigate platform effects using two such datasets, the Netflix Prize dataset and the Facebook New Orleans data, in which we observe seeming discontinuities in user behavior but that we know or suspect are the result of a change in platform design. For the Netflix Prize, we estimate user ratings changing by an average of about 3% after the change, and in Facebook New Orleans, we find that the introduction of the 'People You May Know' feature locally nearly doubled the average number of edges added daily, and increased by 63% the average proportion of triangles created by each new edge. Our work empirically verifies several previously expressed theoretical concerns, and gives insight into the magnitude and variety of platform effects.

Introduction

In social media data, the design and technical features of a given platform constrain, distort, and shape user behavior on that platform, which we call the *platform effects*. For those inside companies, knowing the effect a particular feature has on user behavior is as simple as conducting an A/B test (i.e., a randomized experiment), and indeed such testing is central to creating platforms that shape user behavior in desirable ways. But external researchers have no access to the propriety knowledge of these tests and their outcomes. This is a serious methodological concern when trying to generalize human behavior from social media data: in addition to multiple other concerns, observed behavior could be artifacts of platform design. This concern has thus far only been raised theoretically (Tufekci 2014; Ruths and Pfeffer 2014), and not yet addressed empirically. Even theoretically, the problem is deeper and more subtle than has been appreciated; it is not just a matter of non-embedded researchers having access to the data (Savage and Burrows 2007; Lazer et al. 2009; Huberman 2012; boyd and Crawford 2012), but also that even when researchers have access, without full knowledge of the platform engineering and the decisions and internal research that went into design decisions, the data can be systematically misleading.

One way to study and quantify platform effects as an external researcher is to look for available data that include a significant platform change. Making the assumption that, in absence of the exogenous shock (the change) the previous 'trend' would have remained the same, we can apply the observational inference method of *regression discontinuity design* (Imbens and Lemieux 2008; Lee and Lemieux 2010; Li 2013). While not as certain as experimental design, observational inference methods are the best available way for outside researchers to understand the effects of platform design.

We select two data sets: the Facebook New Orleans data collected by Viswanath et al. (2009), and the Netflix Prize data, described by Koren (2009b). This is no longer publicly available since the close of the Netflix prize, although the terms of use do not mention any expiration on use for those who have already downloaded it.

In the Netflix Prize data set, Koren (2009b), a member of the team that ultimately won the prize (Koren 2009a), points out a curious spike in the average ratings in early 2004. As such a change has modeling implications (previous data should be comparable in order to properly use for training purposes), he explores the possible reasons for this, ultimately identifying an undocumented platform effect as the most likely driver. Then, the Facebook New Orleans data contains an identified, and ideal, example of a platform effect: a clear exogenous shock and a dramatic difference after, through the introduction of the "People You May Know" (PYMK) feature on March 26, 2008. This discontinuity is only mentioned in Zignani et al. (2014); the original paper of the data collectors (Viswanath et al. 2009) does not mention it (although, in another example of a platform effect in collected data, they do note that on July 20, 2008, Facebook launched a new site design that allowed users to "more easily view wall posts through friend feeds" which they use to explain a spike in wall posts towards the end of the collected data).

In sum, we re-analyze the Netflix Prize and Facebook New Orleans data to study possible platform effects in the data. The contributions of this paper are:

- To empirically verify previously expressed theoretical concerns about the possible effects of platform design on the generalizability and external validity of substantive (social scientific) conclusions;
- To import into the social media research community a statistical model that allows quantitative estimation of platform effects;
- To quantify two specific cases of common platform effects, the effect on a social network of a triadic closure-based recommender system and the effect of response item wordings on user ratings.

Background and Related Work

Authors from multiple disciplines (Tufekci 2014; Ruths and Pfeffer 2014) have expressed methodological concerns that the processes found in data derived from social networking sites cannot be generalized beyond their specific platform. Most troublingly, the same things that would cause results to not generalize, such as nonrepresentative samples, idiosyncratic technical constraints on behavior, and partial or uneven data access, are generally unknown and undetectable to an outside researcher (and potentially even to engineers and embedded researchers). Some innovative methods of data comparison have been used to derive demographic information in social media data (Chang et al. 2010; Mislove et al. 2011; Sloan et al. 2013; Hecht and Stephens 2014; Longley, Adnan, and Lansley 2015; Malik et al. 2015) and to identify biases in public APIs (Morstatter et al. 2013; Morstatter, Pfeffer, and Liu 2014), but platform effects remain empirically unaddressed. Part of the problem is that social media platforms are private companies that seek to shape user behavior towards desirable ends, and do so in competition with one another (van Dijck 2013; Gehl 2014); thus, the details of features and functionality which successfully guide user behavior are understandably proprietary in ways that representation and data filtering need not be. The results of research experiments, most notably Kramer, Guillory, and Hancock (2014), deal only indirectly with platform design and engineering. Outside accounting via testing inputs (Diakopoulos 2014) is an important way of identifying overall effective outcomes, but such cross-sectional audits lack a baseline to know how much a given platform design successfully shapes behavior.

Instead, one way to study the problem is the econometrics approach of finding cases that can be treated as 'natural experiments' (Angrist and Pischke 2008; Gelman 2009). We have located two such instances, the Facebook New Orleans data and the Netflix Prize data, where known or suspected change in the platform led to a shift, documented in publicly available data.

Zignani et al. (2014) used the data of the Facebook New Orleans network (Viswanath et al. 2009), along with data from the Chinese social networking site Renren, to investigate the delay between when it is possible for an edge or triangle to form (respectively, when a node enters the network, and when two nodes are unconnected but share a neighbor) and when it actually forms, which they respectively term *link delay* and *triadic closure delay*. They note that on March 26, 2008, there is a drastic increase in the number of links and triangles (our version of those plots given in figs. 1 and 2), corresponding to the introduction of Facebook's "People You May Know" (PYMK) functionality. While this was not the central investigation of their paper, they used it as an opportunity to see how an external feature changed their proposed metrics. They find that this increase consists primarily (60%) of links delayed by over 6 months, and also includes many (20%) links delayed by more than a year. They continue to note, "Although the link delay [metric] reveals interesting characteristic in edge creation process, it is not able to capture the reason behind it, *i.e.* which process causes the observed effects or which algorithms were active in the early rollout of the PYMK feature." However, from their finding that far more triangles were created than edges (based on their fig. 2b, the ratio of new triangles to new edges rose from about 2 before the introduction to about 4 afterwards), it suggests that the created edges were based heavily on triadic closure. They conclude that the external introduction of PYMK manipulated a parameter or parameters of the underlying dynamic network formation process, and furthermore, it did not increase the link creation or triadic closure uniformly, but with bias towards more delayed links and triads. While they say they were able to quantify the effects and impact of the PYMK feature, this did not include estimating the local average treatment effect, which is our specific interest.

The goal of the Netflix Prize competition was prediction and not explanation (Shmueli 2010; Breiman 2001), for which it is not necessary to understand the spike (only to account for it in a model, in order to effectively use past data for training). However, checking for data artifacts is fundamental for any type of data model, and Koren (2009b) devotes some time to investigating an odd spike observed in average ratings in early 2004, about 1500 days into the data set (this plot is recreated in our fig. 3). He proposes and explores three hypotheses:

1. Ongoing improvements in Netflix's 'Cinematch' recommendation technology and/or in the GUI led people to watch movies they liked more;
2. A change in the wordings associated with numerical ratings elicited different ratings (e.g., perhaps a rating of 5 was originally explained as "superb movie" and then was changed to "loved it");
3. There was an influx of new users who on average gave higher ratings.

By noting that the shift also occurs among users who were present both before and after the observed increase, he rejects the third possibility. He finds some support for the first possibility from a model that decomposes ratings into a baseline effect and a user-movie interaction effect (which corresponds to the extent to which users rate movies "suitable for their own tastes"); the interaction effect shows a smooth increase and the baseline has less variability, but there is is still clearly a sudden jump in the baseline. He

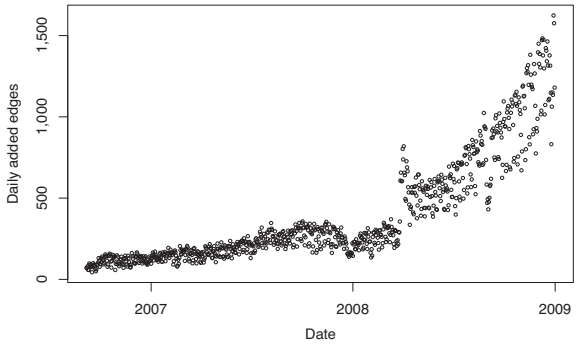

Figure 1: Observed edges added (friendship ties made) in Facebook New Orleans data.

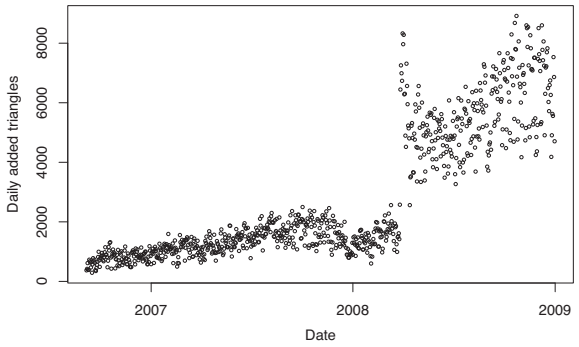

Figure 2: Triangles created with the added edges in Facebook New Orleans data.

writes, "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." Note that the change in wordings associated with numerical ratings is Koren's guess to what the change was; he specifies that uncovering exactly what the "something else" was "may require extra information on the related circumstances." That such a change in wording *could* produce a shift in ratings is supported by decades of research in survey research into response options (Dillman, Smyth, and Christian 2014), but otherwise no further evidence is given.

Data and Methods

Facebook New Orleans

Viswanath et al. (2009) detail how they collected the Facebook New Orleans data through a manual crawl of the New Orleans network, starting from a single user and using breadth-first search. Considering that Facebook started as a college-based network, the boundary specification (Laumann 1973) of users who added themselves to the "New Orleans" network primarily (or those who chose to add it secondarily, perhaps after a college network) may not meaningfully match the college-centric boundaries within which links actually formed (especially since, as the authors point out, regional networks have more lax security than univer-

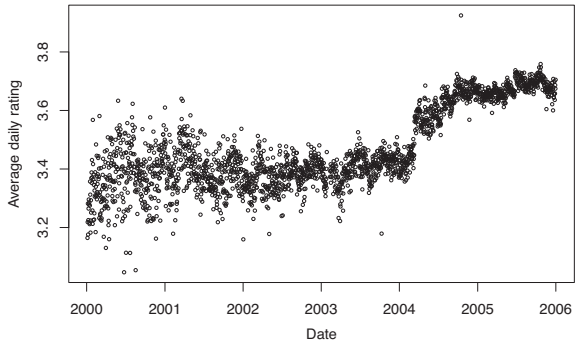

Figure 3: Observed daily averages for the Netflix Prize data.

sity networks, which require a valid email address from the university's domain). Second, only visible profiles could be accessed: the authors estimate, by comparison with statistics from Facebook, that they collected 52% of the users in the New Orleans network.

The Facebook data come in the form of timestamps of added edges. About 41.41% of edges do not have a timestamp. On the data download page, Viswanath et al. (2009) write that "the third column is a UNIX timestamp with the time of link establishment (if it could be determined, otherwise it is [blank])" without elaborating on the reasons for missing labels; we make the assumption that these were the edges already present at the start of data collection. However, we find a great deal of repeated edges. Of the 1,545,686 rows of data, there are only 817,090 unique edges (47.14% of rows are redundant). Breaking it down, of the 640,122 rows that have no timestamp, only 481,368 represent unique edges, and of the 905,564 rows that have a timestamp, only 614,796 represent unique edges. 88,494 edges are repeated twice, 728,596 edges are repeated three times, and no edge is repeated more than three times. We make the decision to drop these repeated edges, assuming that repetition was the result of a repeat visit from multiple crawls (and assuming that timestamps were gathered by the time of detection via BFS, rather than extracted from profiles).

To the unlabeled edges we assign the minimum time present among the remaining edges, and for repeated edges we take their first instance only. Using the igraph library (Csárdi and Nepusz 2006) we take the initial graph and calculate the number of edges, the number of nodes (i.e., non-isolates), the number of triangles, and the transitivity. Since the inter-arrival times are not particularly relevant for our question, we care only about the change in the relative rate over time, we aggregate our analyses by day to create time series: for each day, we add the edges that appeared on that day and recalculate the graph metrics. After, we also calculate the daily density using $2M/(N^2 - N)$ for the number of nodes N and number of edges M. We then difference each of these series, and for each day get the number of edges added, the number of nodes added, the number of new triangles, the change in transitivity, and the change in graph density. (Note that daily aggregation followed by differencing is equivalent to a histogram with day-wide bins, as Zignani et

al. [2014] do for the number of triangles and edges.)

Netflix Prize

The Netflix data come in the form of text files for individual movies, with each line being the rating that a given user gave along with the date from 1999-11-11 to 2005-12-31. Following Koren (2009b)'s plot, we take the daily average in order to see the sudden jump. Examining the number of ratings (i.e., the number of binned observations) per day, we find that they increase linearly in log scale. However, until 1999-12-31, ratings are not daily and even when present are small, whereas from 2000-01-05 (the next day for which there is data) there are daily ratings in the thousands. We take only the data on and after 2000-01-05.

Our own investigation pinpointed the discontinuity as occurring on or around March 12, 2004. We could not find any public record of a platform change at that time nor any clues in press releases around then, and Netflix did not respond to a request for further information.

Statistically, the Netflix data are more straightforward as there is no social network.[1] However, the independence assumptions are more complicated; with a single dynamic network as in the Facebook New Orleans data, we can assume that the network-level rate metrics like the number of added triangles are independent observations across days. If we only consider the average daily rating, we do not take into account multiple ratings by the same individual (and, as Koren [2009b] notes, it is important to correct for different baseline average ratings across users, e.g. making sure an overall 'stingy' user's ratings are comparable to those of an overall 'generous' user). But our interest is not in a full model of user ratings (predictive or explanatory), only a model of the average change to user behavior from a suspected platform effect. That is, we are interested in the marginal effect for which such dependencies are not relevant, and for which we can invoke the random sampling on ratings as a guarantee that our estimate will not have biases in representation.

Causal estimation with discontinuities

Regression discontinuity (RD) design is used to estimate causal effects in cases where there is an arbitrary (and preferably strict) cutoff along one covariate. As shown in Hahn, Todd, and Van der Klaauw (2001), when the appropriate conditions are met, the treatment is effectively random in the left and right neighborhoods of the cutoff c. Causal effects are defined in terms of counterfactuals Y_{0i} (the value of the response were observation i to not be treated) and Y_{1i} (the value of the response were i to be treated); the difference between the two at the time of intervention for treated populations is called the *local average treatment effect* (Imbens and Angrist 1994), α. Given an observed Y_i, this is given by

$$\alpha \equiv E(Y_{1i} - Y_{0i} | x_i = c) = \lim_{x \downarrow c} E(Y_i | x_i = x) - \lim_{x \uparrow c} E(Y_i | x = c) \quad (1)$$

[1] Netflix did briefly attempt to add social networking features in late 2004. However these were discontinued in 2010, with part of the justification being that fewer than 2% of subscribers used the service.

In the linear univariate case, the model is

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 \mathbf{1}(x_i > c) + \beta_3 x_i \mathbf{1}(x_i > c) + \varepsilon_i \quad (2)$$

which effectively fits two separate lines, one for each 'population' before and after the cutoff, with the estimated $\hat{\alpha}$ being the difference between the the two fitted lines at the cutoff. The interest is generally in estimating the causal impact, but as a specification test (Imbens and Lemieux 2008), the joint test for $H_0 : \beta_2 = \beta_3 = 0$ corresponds to a null hypothesis that there is no discontinuity. This model and the corresponding test may be generalized with higher-order polynomial terms. The model also has a natural nonparametric extension: separately fit the same smoother on either side of the discontinuity to estimate the effect, or, test for the discontinuity by seeing if confidence intervals overlap.

Note that the exemplars of RD design are not temporal, and many standard parts of time series modeling are incompatible with RD design. For example, a discontinuity is necessarily nonstationary, and differencing will destroy it (we fitted ARIMA models, and found that differencing was indeed necessary), and similarly, a one-sided moving average smoother applied to both sides of the discontinuity will leave a gap. We found two alternative methodologies created specifically around time series, 'interrupted time series analysis' (McDowall et al. 1980; Wagner et al. 2002; Taljaard et al. 2014) and 'event studies' (MacKinlay 1997), but both are essentially less formal versions of RD design and still neither account for temporal features (namely, autocorrelation). We also tried Gaussian Process (GP) regression (Rasmussen and Williams 2005; MacDonald, Ranjan, and Chipman 2015), as it is able to capture temporal dependencies (Roberts et al. 2012). A squared exponential covariance function gave largely similar results, including posterior intervals about as wide as confidence intervals from other methods (and thus perhaps still not capturing autocorrelation) when fitting separately to either side of the discontinuity. We note that it may be possible in future work to adapt covariance functions that account for 'changepoints' (Garnett et al. 2010) not just to make predictions in the presence of discontinuities, but to do causal inference within the RD framework.

As we are interested in the central tendency rather than on features of the time series, we prioritize the use of the RD framework over time series modeling. To apply RD design, we make the assumption that the respective times at which People You May Know and whatever change took place in Netflix were introduced were effectively random. We use time as the covariate, with the respective cutoffs for the two data sets of 2008-03-26 and 2004-03-12 (i.e., we code for the potential discontinuities starting on those days). We apply both parametric and nonparametric models; for nonparametrics, we use local linear regression as is standard in regression discontinuity design (Imbens and Lemieux 2008) and is also appropriate for time series (Shumway and Stoffer 2011).

While a nonparametric smoother has the advantages of being able to fit cyclic behavior without including specific cyclic terms, confidence intervals still fail to capture the extent of cyclic variance and so are too optimistic even be-

yond not accounting for temporal autocorrelation (Hyndman et al. 2002). Prediction intervals are an alternative as they include the overall variance, but are not straightforward to calculate for smoothers (or generalized linear models, which we use for the count data of daily added edges). Thus, in cases where we can use linear models, we prefer those. Another alternative, which we use for the Netflix data and for edge counts in the Facebook data, is to use local linear quantile regression (Koenker 2005) to get tolerance (empirical coverage) intervals, and specifically, using the interval between a fit to 5% and to 95% to get a 90% tolerance interval (we found too much noise for fits at 97.5% and 2.5% to use a 95% tolerance interval). For consistency, when we do this we also use quantile regression for the central tendency (which corresponds to using the median instead of the mean), which also has the advantage of being more robust to outliers.

Results and Discussion

Netflix Prize data

First, we note that the number of daily ratings increases over time (fig. 4), which corresponds to decreasing variance in the time series plot, suggesting use of weighted least squares. Weighting by the number of daily ratings (so that the days with more ratings are counted more heavily) improved diagnostics across the parametric models we considered; however, we found that the addition of polynomial terms up to and even past 7th order continued to be significant, leading us to prefer the nonparametric approach that can capture the cycles without becoming cumbersome. In fig. (5), we show the results of the local linear quantile regression. As we can see, at the cutoff the two 90% tolerance intervals do not overlap, allowing us to reject the null hypothesis that there is no discontinuity at the 0.10 level.

To test if the model detects jumps at non-discontinuity points, we tried each day as a cutoff. Other than our actual discontinuity, the only points where the tolerance intervals did not overlap were two points before the cutoff we used (March 10th and 11th) and one day after (March 13th). Since we had initially located this date through manual (graphical) investigation, and the choice was not unambiguous within several days, it is unsurprising that the model picks this up as well. While this ambiguity is likely a matter of noise, platform engineers commonly deploy new features gradually to decrease risk, so it is also possible that the ambiguity is a gradual rollout that the model is also detecting.

Sensitivity to the smoothing bandwidth (the tuning parameter which controls the size of the neighborhood used in local fitting) is a concern for estimating the causal effect, so as is recommended, we report the estimates across multiple bandwidths. From 5-fold cross-validation, the optimal bandwidth is 6 (i.e., using kernel $K(x^*, x_i) = \exp\{-.5((x^* - x_i)/6)^2\}$), performed poorly under specification testing, identifying many discontinuities. Larger bandwidths (where the estimator tends towards linear) performed better, but at large bandwidths, again many discontinuities were identified. This is not ideal but unsurprising given the loss function used in quantile regression; quantiles are less swayed by extreme values, such that the non-overlap of tolerance intervals properly capture that there is a discontinuity even far from the actual discontinuity. The estimate of the causal effect may still be good, but with the failure of the specification testing at both low and high bandwidths, we report only within the range that performed well.

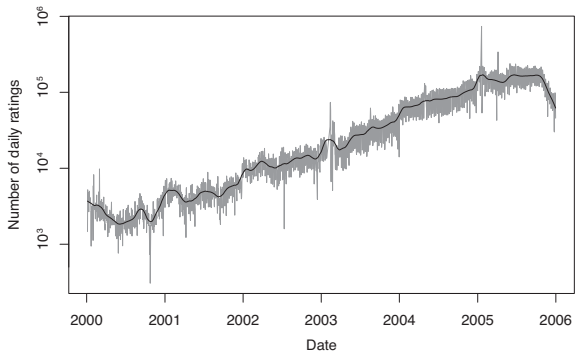

Figure 4: The number of Netflix ratings increases over time (y-axis shown in log scale); and, we can observe from fig. (3), the variance decreases over time, suggesting using the counts as weights. The fitted local linear smoother, which we used for weights, is shown in black. The bandwidth of .026 was selected via 5-fold cross validation.

We estimate the local average treatment effect, the average amount by which the platform change resulted in a change in user ratings, as 0.118 from a bandwidth of 25 (pictured in fig. 5), 0.126 from a bandwidth of 50, 0.124 for a bandwidth of 75, and 0.119 for a bandwidth of 100. Considering the ratings prior to the cutoff had a mean of around 3.44, these amounts are a substantial increase, and are about 3% of the total possible range of ratings (from 1 to 5). This is a less involved case than Facebook, since movie preferences are a relatively low stakes phenomenon, but it shows the application of regression discontinuity. If the cause of the discontinuity is indeed a change in wordings, it shows that, just as in survey research, a change to the format changes the distribution of answers; but unlike in surveys, with large-scale online (streaming) systems, changes become visible as discontinuities in time.

Facebook New Orleans data

Fig. (6) shows the discontinuity in the Facebook New Orleans data across four graph metrics. In addition to the daily counts of the number of added edges and added triangles as examined by Zignani et al. (2014), the discontinuity is pronounced in the transitivity and the density as well (although the units of these are so small as to not be particularly interpretable, so we do not estimate a local average treatment effect).

For the number of edges, we first used a fifth-order polynomial Poisson regression (not pictured), which had excellent regression diagnostics, from which we estimated a local average treatment effect of 356. This is more than a doubling of the pre-cutoff daily average of 314. However, the confidence intervals from the Poisson regression were very

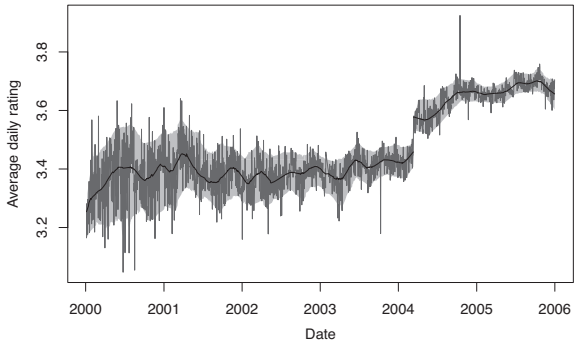

Figure 5: The solid line shows the local linear fit for the median Netflix ratings. The shaded regions are a fitted 90% tolerance interval, from local linear quantile fits to 5% and 95%. The intervals on both sides of cutoff do not overlap.

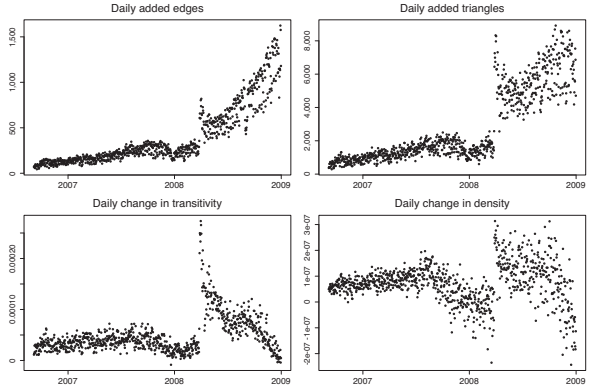

Figure 6: For the Facebook New Orleans data, the daily added edges and triangles created (top left and right, respectively), and the daily change in transitivity and graph density (bottom left and right, respectively).

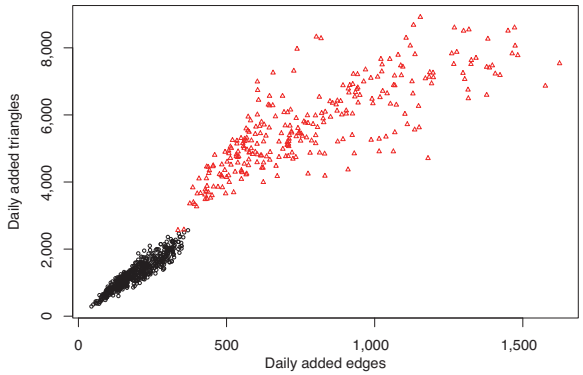

Figure 7: The daily added edges and triangles have a close relationship in the Facebook data. Black circles are time points before 2008-03-26, and red triangles are time points afterwards.

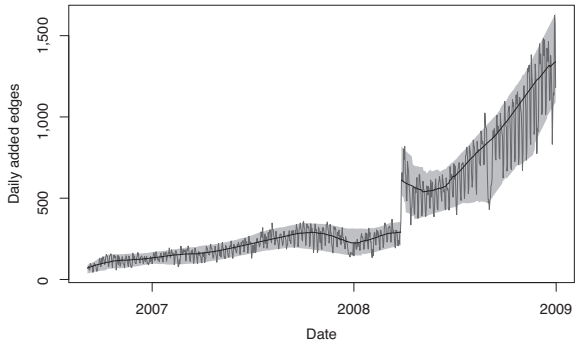

Figure 8: A local linear fit for the median number of edges added daily in Facebook New Orleans. The shaded regions are a fitted 90% tolerance interval, from local linear quantile fits to 5% and 95%. The intervals on both sides of cutoff do not overlap.

narrow and performed poorly under specification testing (as did bootstrap prediction intervals, which were very wide), so we also made fitted tolerance intervals using local linear quantile regression as with the Netflix data, shown in fig. (8). Again, the optimal bandwidth found from 5-fold cross-validation was small and performed poorly under specification testing, as did large bandwidths (tending towards linear). Reporting within the range that performed well under testing, we estimate the local average treatment effect as 319 from a bandwidth of 25 (pictured), 278 for a bandwidth of 50, 228 for a bandwidth of 75, and 201 for a bandwidth of 100.

As the number of edges and triangles are closely related (fig. 7), we follow Zignani et al. (2014) in taking the ratio of triangles to edges. This represents the average number of triangles created by each added edge, and captures the extent of triadic closure on a scale more interpretable scale than that of changes in transitivity (which are in the ten thousandths). For a parametric model with an indicator for the discontinuity as described in eqn. (2), up to fourth-order polynomial terms were significant additions to the model in partial F tests. The fit is shown in fig. (9), which estimates a local average treatment effect of 3.85. This is even more dramatic than the effect in Netflix; given that the mean ratio was estimated at 6.03 before the jump, this is an increase of 63.8%. For specification testing, when fitting curves separately to either side for each timepoint, the prediction intervals are disjoint only at seven consecutive times (i.e., seven points could be considered discontinuities); from two days before the date of the introduction of the PYMK feature to four days after, which we can attribute to the magnitude of the discontinuity.

Conclusion

For much of data analysis, discontinuities (such as from abrupt platform changes in social media) are seen as incidental, or annoyances to be corrected (Roggero 2012). Indeed, they appear in the literature as curiosities or asides. However, given the theoretical concerns about the nature of

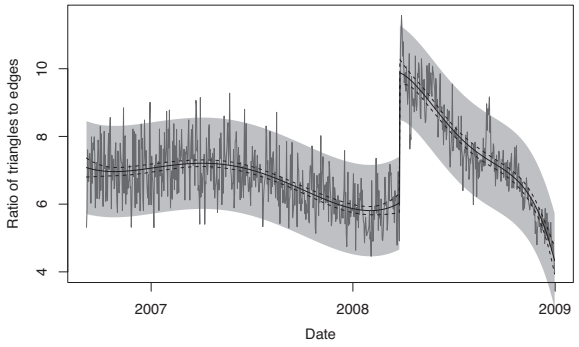

Figure 9: A fourth-order polynomial regression model fitted to the daily ratio of added triangles to added edges in Facebook New Orleans. The solid line is the fit, the dashed line is a 95% confidence interval, and the shaded region is a 95% prediction interval.

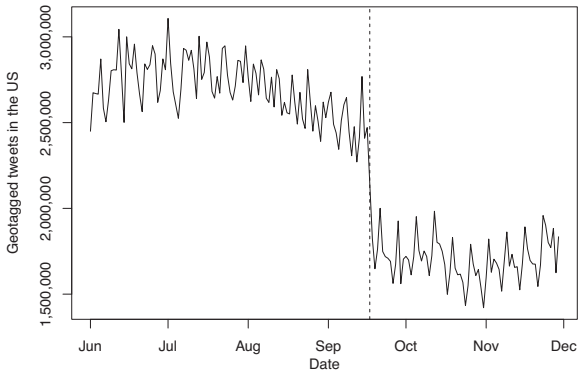

Figure 10: Another potential discontinuity, seen amidst cyclic behavior in the volume of geotagged tweets collected in the US in 2014.

social media data, they can give valuable insights. Our finding about the 3% change in average Netflix ratings echoes work in survey research about response item wordings in a different setting and with different sort of data, quantifying how much we might expect a platform change to shift a baseline. For the Facebook New Orleans data, the finding is even more dramatic and widely applicable: we now have a sense that the introduction of a triadic closure-based recommender system can nearly double the rate of link creation. Furthermore, it changes the nature of the created links (focusing on closing triads), which has repercussions for the graph structure, seen for example in the changes in density. This provides an empirical extension of a concern raised by Schoenebeck (2013) about how variation in technology adoption creates online social networks that differ systematically from the underlying social network: from our results, we see it is not just the process of joining social networking sites that creates observed network properties, but also the ways in which platforms design influences users. Multiple works have considered whether network metrics of large online social networks differ from those of previously studied social networks (Corten 2012; Quercia, Capra, and Crowcroft 2012; Ugander et al. 2011; Mislove et al. 2007); we can continue to theorize how differences result from platform effects, usage patterns, and demographic representation, rather than from online platforms being a superior way to measure social networks.

There are concerns about what social media ties even represent (Lewis et al. 2008), with some authors pointing to interactions over ties (Viswanath et al. 2009; Romero et al. 2011; Wilson et al. 2012; Jones et al. 2013) as more meaningful than the existence of ties. But our results show that the problem is not just one of ties not being a rich enough measure, but that they a non-naturalistic measure of social relationships, and furthermore, their existence determines visibility and access and thereby what activity happens. As people accept suggested links and begin interacting, the underlying phenomenon (the relationships and the network effects) changes, whether for good (Burke and Kraut 2014) or ill (Kwan and Skoric 2013). On Netflix, if changes affect different movies differently, it has consequences for modeling user behavior preferences. Beyond research concerns, there are economic benefits for the creators of movies that benefit from platform changes. Lotan (2015) observed this potentially happening in Apple's App Store, where what appeared to be an (unannounced, undocumented) engineering change in the search results ranking led to changes in app sales.

Regression discontinuity design has a rich literature, and there are likely many other case to which to apply it in social media data. For example, Li (2013) identified Yelp ratings being rounded to the nearest star as an opportunity to apply RD design. As another example, in geotags we collected from the US in 2014, there was a sudden decrease (fig. 10) on September 18th, the same day Twitter released significant updates to profiles on Twitter from iPhone.[2] Other such examples can allow outside researchers to start building up a public body of knowledge about the ways in which platform design are responsible for observed behavior. Extensions to regression discontinuity are also relevant, for example in how Porter and Yu (2015) develop specification tests into tests for unknown discontinuities.

Social media data has been compared to the microscope in potentially heralding a revolution in social science akin to that following the microscope in biology (Golder and Macy 2012). This metaphor may have a deeper lesson in a way that its advocates did not expect: history of science has shown (Szekely 2011) that it was not a simple process to connect the new instrument, with its multiple shortcomings, to the natural objects it was supposedly being used to study. It took centuries of researchers living with the microscope, improving the instrument but also understanding how to use it (e.g., recognizing the need for staining, or the importance of proper lighting), that microscopes became a firm part of rigorous, cumulative scientific research. We would hope that social media data will not take as long, but at the same time,

[2]"A new profile experience on Twitter for iPhone," September 18, 2014, https://blog.twitter.com/2014/a-new-profile-experience-on-twitter-for-iphone, accessed 1/2016.

it is as necessary as ever to question the relationship between the novel instrument and the object of study.

Acknowledgements

Momin is supported by an award from the ARCS Foundation.

References

Angrist, J. D., and Pischke, J.-S. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

boyd, d., and Crawford, K. 2012. Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15(5):662–679.

Breiman, L. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science* 16(3):199–231.

Burke, M., and Kraut, R. E. 2014. Growing closer on Facebook: Changes in tie strength through social network site use. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, 4187–4196.

Chang, J.; Rosenn, I.; Backstrom, L.; and Marlow, C. 2010. ePluribus: Ethnicity on social networks. In *Proceedings of the Fourth International Conference on Weblogs and Social Media*, ICWSM-10, 18–25.

Corten, R. 2012. Composition and structure of a large online social network in the Netherlands. *PLoS ONE* 7(4):e34760.

Csárdi, G., and Nepusz, T. 2006. The igraph software package for complex network research. *InterJournal* Complex Systems(1695).

Diakopoulos, N. 2014. Algorithmic accountability reporting: On the investigation of black boxes. Tow Center for Digital Journalism, Columbia Journalism School.

Dillman, D. A.; Smyth, J. D.; and Christian, L. M. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. John Wiley & Sons, Inc., 4th edition.

Garnett, R.; Osborne, M. A.; Reece, S.; Rogers, A.; and Roberts, S. J. 2010. Sequential Bayesian prediction in the presence of changepoints and faults. *The Computer Journal* 53(9):1430–1446.

Gehl, R. W. 2014. *Reverse Engineering Social Media: Software, Culture, and Political Economy in New Media Capitalism*. Temple University Press.

Gelman, A. 2009. A statistician's perspective on "Mostly Harmless Econometrics: An Empiricist's Companion", by Joshua D. Angrist and Jörn-Steffen Pischke. *Stata Journal* 9(2):315–320.

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

Hahn, J.; Todd, P.; and Van der Klaauw, W. 2001. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69(1):201–209.

Hecht, B., and Stephens, M. 2014. A tale of cities: Urban biases in volunteered geographic information. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, ICWSM-14, 197–205.

Huberman, B. A. 2012. Sociology of science: Big data deserve a bigger audience. *Nature* 482(7385):308.

Hyndman, R. J.; Koehler, A. B.; Snyder, R. D.; and Grose, S. 2002. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting* 18(3):439–454.

Imbens, G. W., and Angrist, J. D. 1994. Identification and estimation of local average treatment effects. *Econometrica* 62(2):467–475.

Imbens, G. W., and Lemieux, T. 2008. Regression discontinuity designs: A guide to practice. *Journal of Econometrics* 142(2):615–635.

Jones, J. J.; Settle, J. E.; Bond, R. M.; Fariss, C. J.; Marlow, C.; and Fowler, J. H. 2013. Inferring tie strength from online directed behavior. *PLoS ONE* 8(1):e52168.

Koenker, R. 2005. *Quantile Regression (Econometric Society Monographs)*. Cambridge University Press.

Koren, Y. 2009a. The BellKor solution to the Netflix Grand Prize.

Koren, Y. 2009b. Collaborative filtering with temporal dynamics. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, 447–456.

Kramer, A. D. I.; Guillory, J. E.; and Hancock, J. T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* 111(24):8788–8790.

Kwan, G. C. E., and Skoric, M. M. 2013. Facebook bullying: An extension of battles in school. *Computers in Human Behavior* 29(1):16–25. Including Special Section Youth, Internet, and Wellbeing.

Laumann, E. O. 1973. *Bonds of pluralism: The form and substance of urban social networks*. New York: John Wiley & Sons, Inc.

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.; and Van Alstyne, M. 2009. Computational social science. *Science* 323(5915):721–723.

Lee, D. S., and Lemieux, T. 2010. Regression discontinuity designs in economics. *Journal of Economic Literature* 48(2):281–355.

Lewis, K.; Kaufman, J.; Gonzalez, M.; Wimmer, A.; and Christakis, N. 2008. Tastes, ties, and time: A new social network dataset using Facebook.com. *Social Networks* 30(4):330–342.

Li, X. 2013. How does online reputation affect social media endorsements and product sales? evidence from regression discontinuity design. In *The 24th Workshop on Information Systems Economics*, WISE 2013.

Longley, P. A.; Adnan, M.; and Lansley, G. 2015. The geotemporal demographics of Twitter usage. *Environment and Planning A* 47(2):465–484.

Lotan, G. 2015. Apple's App charts: 2015 data and trends ...or how much harder it is to get into the top charts. *Medium*, 15 December 2015. Available at https://medium.com/i-data/apple-s-app-charts-2015-data-and-trends-abb95300df57.

MacDonald, B.; Ranjan, P.; and Chipman, H. 2015. GPfit: An R package for fitting a gaussian process model to deterministic simulator outputs. *Journal of Statistical Software* 64(12):1–23.

MacKinlay, A. C. 1997. Event studies in economics and finance. *Journal of Economic Literature* 35:13–39.

Malik, M.; Lamba, H.; Nakos, C.; and Pfeffer, J. 2015. Population bias in geotagged tweets. In *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research*, ICWSM-15 SPSM, 18–27.

McDowall, D.; McCleary, R.; Meidinger, E. E.; and Jr., R. A. H. 1980. *Interrupted Time Series Analysis*. Number 21 in Quantitative Applications in the Social Sciences. SAGE Publications, Inc.

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P.; and Bhattacharjee, B. 2007. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, IMC '07, 29–42.

Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. 2011. Understanding the demographics of Twitter users. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, ICWSM-11, 554–557.

Morstatter, F.; Pfeffer, J.; Liu, H.; and Carley, K. 2013. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, ICWSM-13, 400–408.

Morstatter, F.; Pfeffer, J.; and Liu, H. 2014. When is it biased?: Assessing the representativeness of Twitter's streaming API. In *Companion to the Proceedings of the 23rd International Conference on World Wide Web*, WWW Companion '14, 555–556.

Porter, J., and Yu, P. 2015. Regression discontinuity designs with unknown discontinuity points: Testing and estimation. *Journal of Econometrics* 189(1):132–147.

Quercia, D.; Capra, L.; and Crowcroft, J. 2012. The social world of Twitter: Topics, geography, and emotions. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, ICWSM '12, 298–305.

Rasmussen, C. E., and Williams, C. K. I. 2005. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. The MIT Press.

Roberts, S.; Osborne, M.; Ebden, M.; Reece, S.; Gibson, N.; and Aigrain, S. 2012. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 371(1984).

Roggero, M. 2012. Discontinuity detection and removal from data time series. In Sneeuw, N.; Novák, P.; Crespi, M.; and Sansò, F., eds., *VII Hotine-Marussi Symposium on Mathematical Geodesy*, volume 137 of *International Association of Geodesy Symposia*, 135–140. Springer Berlin Heidelberg.

Romero, D.; Meeder, B.; Barash, V.; and Kleinberg, J. 2011. Maintaining ties on social media sites: The competing effects of balance, exchange, and betweenness. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, ICWSM '11, 606–609.

Ruths, D., and Pfeffer, J. 2014. Social media for large studies of behavior. *Science* 346(6213):1063–1064.

Savage, M., and Burrows, R. 2007. The coming crisis of empirical sociology. *Sociology* 41(5):885–899.

Schoenebeck, G. 2013. Potential networks, contagious communities, and understanding social network structure. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, 1123–1132.

Shmueli, G. 2010. To explain or to predict? *Statistical Science* 25(3):289–310.

Shumway, R. H., and Stoffer, D. S. 2011. *Time series analysis and its applications with R examples*. Number 47 in Springer Texts in Statistics. New York: Springer, 3rd edition.

Sloan, L.; Morgan, J.; Housley, W.; Williams, M.; Edwards, A.; Burnap, P.; and Rana, O. 2013. Knowing the tweeters: Deriving sociologically relevant demographics from Twitter. *Sociological Research Online* 18(3).

Szekely, F. 2011. Unreliable observers, flawed instruments, 'disciplined viewings': Handling specimens in early modern microscopy. *Parergon* 28(1):155–176.

Taljaard, M.; McKenzie, J. E.; Ramsay, C. R.; and Grimshaw, J. M. 2014. The use of segmented regression in analysing interrupted time series studies: An example in pre-hospital ambulance care. *Implementation Science* 9(77).

Tufekci, Z. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, ICWSM-14, 505–514.

Ugander, J.; Karrer, B.; Backstrom, L.; and Marlow, C. 2011. The anatomy of the Facebook social graph.

van Dijck, J. 2013. *The Culture of Connectivity: A Critical History of Social Media*. Oxford University Press.

Viswanath, B.; Mislove, A.; Cha, M.; and Gummadi, K. P. 2009. On the evolution of user interaction in Facebook. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN '09)*, 37–42.

Wagner, A. K.; Soumerai, S. B.; Zhang, F.; and Ross-Degnan, D. 2002. Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics* 27(4):299–309.

Wilson, C.; Sala, A.; Puttaswamy, K. P. N.; and Zhao, B. Y. 2012. Beyond social graphs: User interactions in online social networks and their implications. *ACM Transactions on the Web* 6(4):17:1–17:31.

Zignani, M.; Gaito, S.; Rossi, G. P.; Zhao, X.; Zheng, H.; and Zhao, B. 2014. Link and triadic closure delay: Temporal metrics for social network dynamics. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, ICWSM-14, 564–573.