The Workshops of the Tenth International AAAI Conference on Web and Social Media
News and Public Opinion: Technical Report WS-16-18

Measuring Click and Share Dynamics on Social Media: A Reproducible and Validated Approach

Lucy X. Wang
BuzzFeed
111 E. 18th St.
Fl 16
New York, NY 10003
lucy.wang@columbia.edu

Arthi Ramachandran
Columbia University
500 West 120 Street, Room 450
MC0401
New York, NY 10027
arthir@cs.columbia.edu

Augustin Chaintreau
Columbia University
500 West 120 Street, Room 450
MC0401
New York, NY 10027
augustin@cs.columbia.edu

Abstract

Online social conversations have increasingly become the means to find and view information online. In contrast to traditional web surfing models, clicks from social media result from a series of endorsements subject to user memory, past behavior and intermittently divided attention. Understanding click dynamics allows us to leverage those facets to improve relevance, forecast traffic and better manage influence in information dissemination. Unfortunately, data on clicks – even in aggregate – remain proprietary and inaccessible to researchers and scientists in many disciplines.

In our work, we aim to allow the study of clicks through a proxy, allowing analysts to fully study click dynamics on Twitter. We focus on the scope of a news content publisher with a large readership and a broad domain of topics. We validate one such proxy, clicks-per-follower (CPF), based on publicly accessible data. We develop a model to compute CPI from public data. We use this method to examine how sharing affects consumption on Twitter: our findings suggest that mass retweeting of a URL does not necessarily translate into a substantial increase in clicks.

1 Introduction

Social media provides an opportunity to observe individual and collective influences at work. In today's web, one out of three visits gets triggered by a social media action (*e.g.*, a share, a retweet, a pin), but those clicks remain a sparse and tiny fraction of content each online user selects from a torrent of information directed at her feeds (Wong 2015). How users select what information to read, and what it tells us about influence, today, appears out of the reach of researchers and social scientists. Observing the click dynamics in relation to online conversation could lead to multiple new vantage points on user behaviors and algorithms leveraging those to improve relevance, forecast traffic, and more generally manage influence in information dissemination.

In this paper, we offer a unique comparison of click-focused metrics for social media conversation. Our goal is to unlock the potential of studying social media consumption dynamics.

- First, we show that studying social clicks is technically feasible and accurate using publicly available data. We validate, for the first time, that the Click-Per-Follower (CPF) metric recently introduced can be used as a proxy for understanding reaction to content in a similar manner as the traditional click-through-rate. (Section 2)
- Going further, we present an intuitive linear model that allows one to accurately estimate Click-Per-Impression on Twitter using several publicly available data features, with an R^2 of 0.997. (Section 3)
- Finally, we apply this methodology to highlight how sharing and click dynamics differ on Twitter. We found that high retweeting volume is not necessarily always correlated with more active click engagement. (Section 4)

While we briefly survey the only known work on clicks from social media referrals, we note that no prior work reconciled the view of a publisher accessing performance of its own content (in this case, a major online news source) with the traffic analysis tools available publicly to researchers. Previous observations confirmed the expectation that in general, online user reading behaviors might radically differ from their posting activities, but we are not aware of previous quantitative observations in which endorsements are shown in such a context to have fatiguing effects.

2 Data Collection and Computing Click-Through Rates

In our study, we focus on Twitter data from three major sources: 1) Publisher dataset from Twitter Analytics (private), 2) Retweet dataset from Twitter Rest API (public), and 3) Clicks dataset from Bit.ly's API (public).

2.1 Publisher Dataset

Buzzfeed is an internet media company focussing on creation and distribution of content. They cover a wide range of topics across multiple platforms. On Twitter, they have $\sim$ 40 active posting accounts, each targeted to a different demographic. As a publisher on Twitter, they have access to Twitter Analytics of their account, including the metrics of link clicks, retweets, and impressions. These metrics gives us more granular data of the readership of an account. We leveraged these analytics for Buzzfeed Twitter accounts which include links to `www.buzzfeed.com` content. We focused on original tweets, excluding retweets of

non-BuzzFeed user tweets, in order to preserve uniformity of content source. The readership information provided by this dataset is used solely here for the purpose of validating a reproducible model.

Our dataset includes all tweets published by any of BuzzFeed's Twitter accounts over a 7-day period from February 23 to March 2, 2016 (4K tweets). The largest account is BuzzFeed's primary, eponymous account, which has 2.8M followers. This account posts a wide variety of links to BuzzFeed web articles, typically those projected to become most viral. The next most popular account is BuzzFeedNews, with 470K followers, which posts links to traditional news stories published by BuzzFeed. The remaining accounts serve a more specific niche or content genre, and are named accordingly (e.g. BuzzFeedSports, BuzzFeedFashion).

2.2 Retweet Dataset

We used Twitter's REST API to scrape all tweets published by BuzzFeed accounts and all related public retweets over the span of the same 7 days, forming a complete public dataset of the $\sim$ 4K unique BuzzFeed tweets. Each (re)tweet also provides the publicly available follower count of the (re)tweeting user.

2.3 Clicks Dataset

To further supplement our public readership data, we used bit.ly's API to gather all twitter-originating link clicks (those with `twitter.com` or `t.co` as the listed referrer domain). Of the 6K tweets from a 14-day range (Feb 14-Mar 2, 2016), we considered only those 2.4K tweets with bit.ly URLs. This was the sole dataset used for our click analyses. Note that most Buzzfeed accounts almost exclusively used one method of sharing information (either using bit.ly for almost all the links they post, or not at all). As a result, focusing on bit.ly links introduces a source-bias since accounts behave differently. However, we believe there is little intrinsic bias introduced by bit.ly itself.

2.4 Click-Through-Rate Made Accessible

The click through rate (CTR) of content is a very valuable piece of information. In practice, however, it is often challenging to compute without access to good quality data regarding the readership. We compute two types of CTR:

- Clicks Per Impression (CPI): $\frac{\#\ clicks}{\#\ impressions}$
- Clicks Per Follower (CPF): $\frac{\#\ clicks}{\sum_{u \in U} \#followers(u)}$ where $U =$ the set of users tweeting or retweeting the link

The main difference between the two metrics is how we compute the audience size. With CPI, we consider the audience as the number of Twitter users who have been exposed to the URL, or the number of impressions. While this is an accurate measurement of CTR, it is often hard to measure with public data. In contrast, for CPF, we consider the audience to be the sum of the number of followers for every Twitter user who tweeted or retweeted the given url. This method can overestimate the number of impressions and capture too much noise since total follower counts fail to account for 1) the overlap of follower sets and 2) the level of activity of followers. While one can theoretically compute the number of unique followers to account for overlap, the number of api queries involved quickly makes this prohibitively expensive. Previous work quantified this overestimation from overlap, finding it is less than 20% for 75% of follower counts (Gabielkov et al. 2016).

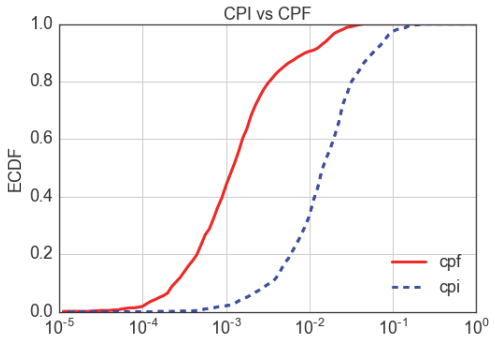


Figure 1: Clicks per Follower (Red) and Clicks per Impression (Blue): CPI is computed from the publisher dataset. CPF is computed entirely from public data.

2.5 Clicks Per Follower versus Clicks Per Impression

Figure 1 compares the ecdf distributions of CPI and CPF for each URL. Note that CPI is computed from our publisher dataset and CPF is computed entirely from public data: the number of clicks from bit.ly, and the number of followers from Twitter. While the magnitudes of CPI and CPF differ by a factor, they follow the same general trend. In the next section we develop a means to use CPF as a proxy for CPI.

3 Using CPF as a proxy for CPI

We know from previous studies that CPI estimations are skewed by estimation errors in both clicks and impressions. We develop and validate a means to estimate CPI based on bit.ly click data. In addition to previously known factors resulting in overestimation of impressions, we find that there are factors that result in underestimating the number of impressions - namely, the existence of private users.

3.1 Estimating Clicks

With the public availability of bit.ly link click data, it is relatively easy for non-publishers to track engagement. When we compare the distributions of the actual clicks to the bit.ly clicks, we find that their distributions are almost entirely aligned. Individual tweets are significantly correlated (Figure 3a) (Pearson coefficient, $r = 0.819$, p-value$= 0.0$). All these measurements depend on timing – if data isn't collected at exactly the same time, then there can be small issues that arise. Overall, we see that such timing issues often cause bit.ly clicks to underestimate, but in each case, the effect is quite small. We also observe the opposite effect, where the actual clicks are less than the observed bit.ly

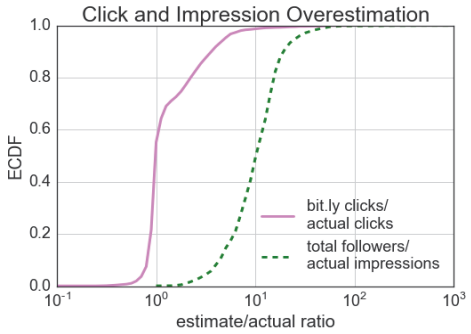


Figure 2: Estimation Errors in Clicks and Impressions

clicks. This can occur when users copy the raw bit.ly link and tweet it, avoiding the retweet button directly on the tweet. Twitter Analytics no longer captures those clicks because they are seen to be part of a different diffusion tree.

We further compare, for each URL, the difference between the bitly clicks and the actual clicks to identify the estimation error. Figure 2 shows the distribution of estimation ratios defined as $\frac{bit.lyclicks}{actualclicks}$. About 51% of links had underestimated clicks, with a mean ratio of 0.87 bit.ly clicks to actual clicks. Among the overestimated clicks, the mean ratio of bit.ly clicks to actual clicks was 3.1. The magnitude of the ratio in overestimation far exceeds that of underestimation, making the mean estimation ratio 11.9.

We developed a linear model for predicting the actual clicks from the bit.ly clicks (Figure 3). We trained on a random 96% of the data. We used bit.ly clicks (BT) and the number of followers of the primary account (PF) as independent variables. Our regression model was:

$$\log(Clicks^*) = \beta_0 + \beta_1 \log(BT) + \beta_2 \log(PF).$$

Figures 3b and 3c show the quality of the regression by overlaying the predicted click values with the actual clicks. The best fit model included both these variable and no intercept (Table 3d). The training fit had a R^2 value of 0.988 while the test prediction had an R^2 of 0.932.

3.2 Over-estimating Impressions - Limitations of Follower Counts

A key question requires quantifying the degree to which we over-estimate the number of impressions. We consider the ratio of total follower count to actual impression count: $\frac{totalfollowers}{impressions}$. We looked at the distribution of the overestimates (Figure 2). Overall, we find that most tweets are overestimated by a relatively small fraction. For $\sim 91\%$ of these URLs, tweets were overestimated by a factor of $0 - 20$.

3.3 Under-estimating Impressions - Limitations of Private Tweets

Twitter has protections for users who opt for increased privacy. In these cases, the Twitter API only allows public access to retweet data on public retweets. In our publisher dataset, we know the total number of retweets and impressions for every tweet. However, when estimating tweet impressions, we are precluded from computing the total number of followers potentially exposed to the tweet, as we can not obtain the follower counts of private retweeters. Given that we know the actual number of total retweets received on every URL from our publisher dataset, we can compute the difference between this total and the number of retweets counted from our public dataset to yield the number of retweets that are private.

Figure 4 shows the cumulative distribution of the underestimation of our retweets. For about 75% of tweets, less than half the retweets are private. On average, 35% of retweets are private. The number of private retweets on a URL and the number of over-counted estimated impressions have a negative relationship: the more private retweets exist for a URL tweet, the smaller the impression over-estimation (Pearson's $r = -0.420$, p-value$= 8.7e^{-}122$.

We thus see that there are two opposing factors in our estimation of impressions from total follower count: 1) overestimation from double-counting followers in overlapping follower sets and counting inactive followers, and 2) underestimation from not counting private retweeter follower sets.

3.4 Estimating Impressions from Followers

Ideally, we want to be able to develop a model of estimation that is robust to both these under- and over- estimate issues. To better understand the relationship between CPF and CPI, we examine how the CPI and CPF correlate for each URL. This correlation is, in fact, quite strong - Pearson's $r = 0.946$ with a very low p value of 0.0 (Figure 5).

We found that, in addition to total follower count (TF), the number of followers of the original poster (primary followers or PF), and the number of non-private retweets (RT) also correlate linearly with impressions. We fitted a multi-variate least squares regression model:

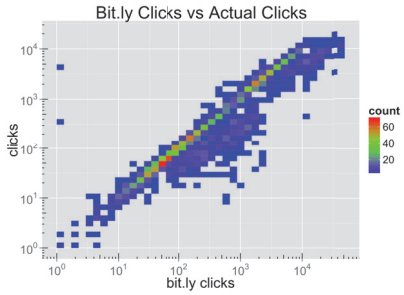
$$\log(Impr^*) = \beta_0 + \beta_1 \log(TF) + \beta_2 \log(PF) + \beta_3 \log(RT)$$

We evaluated the model's fit on the entire dataset and the model's predictive performance when trained on 94% of the dataset (Figure 6d). We considered several models: (i) with just Total Followers, (ii) with both TF and PF, and (iii) with TF, PF, RT, and no intercept. While total followers suffices in estimating impressions, the addition of primary followers boosts the quality of the regression. Including retweets allowed it to perform even better. Given the regression coefficients, it seems that the total number of followers provides crucial information to derive the number of impressions, while both the number of primary followers and the number of retweets help refine the prediction.

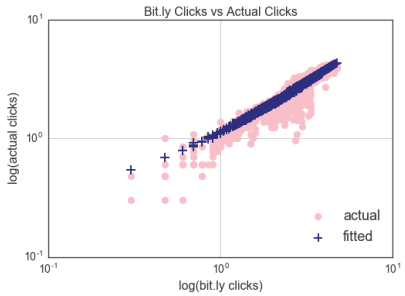
Figures 6a, 6b and 6c show estimated impressions overlaid on the real values for each of the independent variables. In the remaining analyses, we use the estimated impression value derived from: $\log(Impr^*) = 0.7396 \log(TF) + 0.0473 \log(PF) + 0.1027 \log(RT)$.

4 The Effects of Retweeting on Clicks

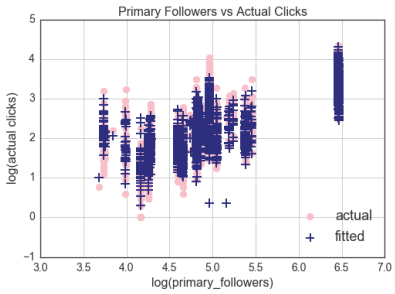
Click rate metrics give an overall view into the performance of a URL. However they miss insight into the cause of the readership of content - its diffusion and sharing characteristics. We would expect that retweeting features bear a



(a) Correlation of Bit.ly Clicks vs Actual Clicks



(b) Clicks* vs Bit.ly clicks



(c) Clicks* vs Primary Followers

Model		Bit.ly Clicks (β_1)	Primary Fol. (β_2)	Intercept (β_0)	R^2
Bit.ly Clicks	**Training**	0.9098	0	0.1055	0.888
Only	**Prediction**	0.9068	0	0.1117	0.918
Bit.ly Clicks + Primary	**Training**	0.8145	0.0711	0	0.988
Followers- Intercept	**Prediction**	0.8113	0.0727	0	0.932

(d) Predicting Clicks: Regressor Coefficients and R^2 Score.

Figure 3: Correlation and Linear Regression for Predicting Clicks from Bit.ly clicks and Primary Followers.

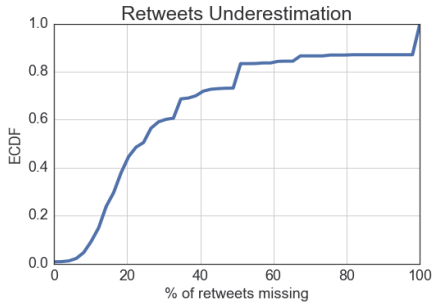


Figure 4: ECDF of Percent of Retweets that are Missing

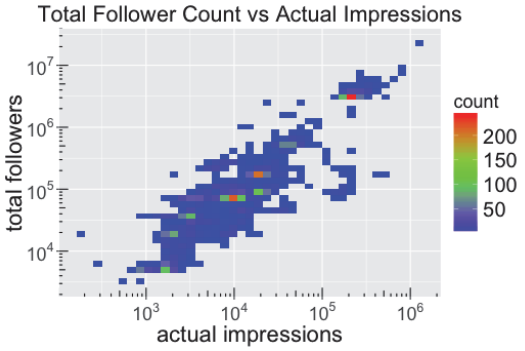


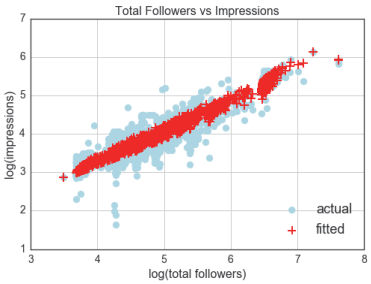
Figure 5: Real Impressions vs Total Follower Count

strong relationship to the audience and the eventual readership since the act of retweeting is the primary mechanism to generate an audience. This type of analysis is relatively new with the work of (Gabielkov et al. 2016) being the first to examine this relationship. They found that there was a strong positive correlation between the number of retweets of a link and the number of clicks. We expand on their work by examining the relationship in content of different audience sizes with a broader set of type of content, rather than traditional news alone.

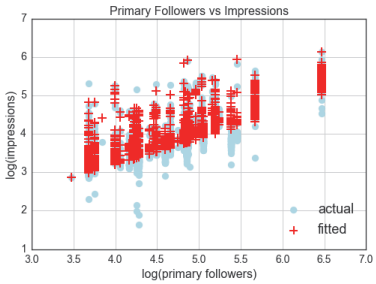
In these analyses, we use publicly attained retweets, estimated impressions (impressions*), estimated clicks (clicks*), and the estimated clicks per impression (CPI*) to analyze the relation of clicks and click rate with share rate.

We first found that the relationship between retweeting and CPI demonstrates a law of diminishing returns of clicks, as it has a slight negative correlation (Pearson's $r = -0.089$, p value$= 1.70e^{-5}$). However, looking into the relationship between retweets and absolute number of clicks presents another picture of the effect of endorsements on news item reach on Twitter. Here the results suggests a law of *no* returns. While this limit on reach has been previously observed in social media (Myers, Zhu, and Leskovec 2012), those studies are based in a different setting, and define reach by re-shares rather than clicks.

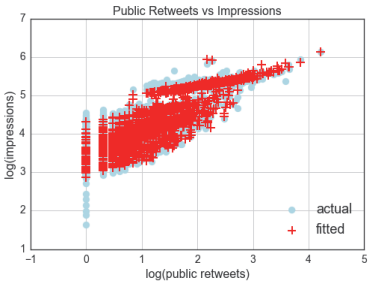
We see a threshold effect at ~ 100 retweets, above which the clicking and sharing relationship changes. When number of retweets is < 100 (Figure 7a), clicks and shares are positively correlated (Pearson's $r = 0.530$). However, past this threshold, increasing retweets does not translate to increasing clicks (Figure 7b)(Pearson's $r = 0.060$, p value$= 0.242$). A diminishing impression rate could explain this exhaustion of reach, but not entirely. While the growth rate of impressions diminishes, impressions do still



(a) Impressions* vs Total Followers



(b) Impressions* vs Primary Followers



(c) Impressions* vs Retweets

Model		Total Fol. (β_1)	Primary Fol. (β_2)	Retweets(β_3)	Intercept (β_0)	R^2
Total Followers Only	**Training**	0.8181	0	0	-0.0661	0.890
	Prediction	0.8180	0	0	-0.0651	0.892
Total Followers + Primary Followers	**Training**	0.8203	-0.0024	0	-0.0654	0.890
	Prediction	0.8187	-0.0012	0	-0.0629	0.905
Total Followers + Primary Followers + Retweets - Intercept	**Training**	0.7421	0.0447	0.1032	0	0.997
	Prediction	0.7396	0.0473	0.1027	0	0.912

(d) Predicting Impressions: Regressor Coefficients and R^2 Score

Figure 6: OLS Linear Regression for Predicting Impressions from Total Followers, Primary Followers, and Retweets.

increase with sharing with a Pearson's $r = 0.603$ (p value = $4.71e^{-39}$) (Figure 7b). Unlike clicks, we don't yet observe a limit to the growth of impressions in our scope of sharing magnitude. We pose three potential future directions:

1. The passage of time that comes with increased retweet counts presents an additional variable of note. To what extent does elapsed time affect the *clickability* of news links, even as these links get increasingly shared?
2. We should not discount the occurrence of retweeting a link without ever clicking on it. Clickability and shareability are two content features which may each encompass very distinct sets of criteria. For example, a headline containing the entire summary of an objective news story may be shared without being clicked. What types of content over-index for shares yet under-index for clicks, and vice versa? How frequently does each scenario occur, and at what magnitude do they over and under index?
3. The phenomena we observe are confounded to some extent by Twitter's own social structures, nature of its activity, and newsfeed algorithms. Do we observe similar reach thresholding in other social networks, like Facebook?

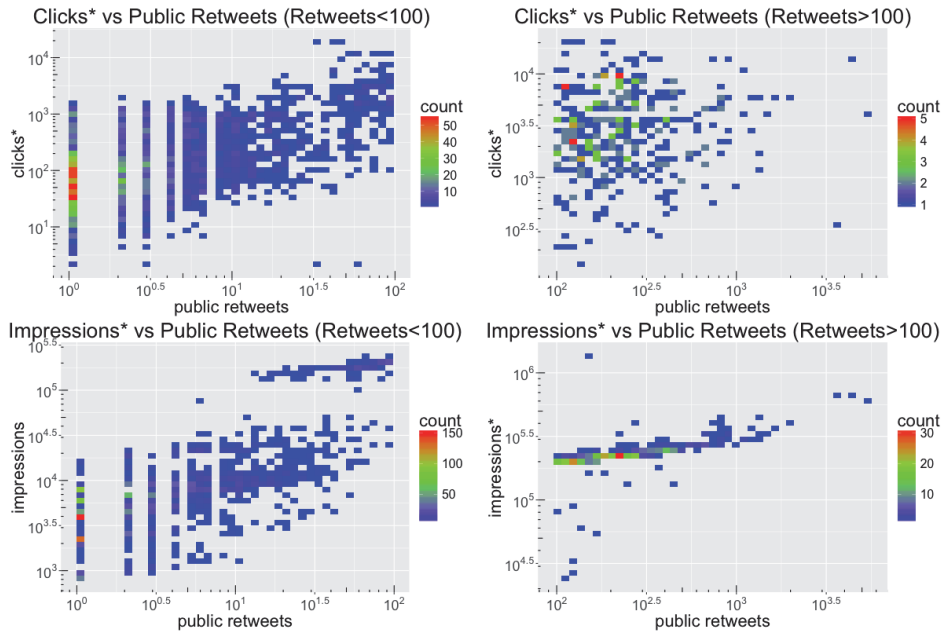
5 Related Work

Motivated by revenue models based on cost-per-clicks (cpc), most prior studies of online clicking habits are specifically targeted at online advertising. Models attempt at measuring the quality of an ads, and the relevance of personalization using its Click-Through-Rate (CTR) a metric resembling CPI in our context (Farahat and Bailey 2012; McMahan et al. 2013). Others attempt at exploiting anomalous behaviors to fight click-fraud (Dave, Guha, and Zhang 2012). In the context of an online social network, (Saez-Trumper et al. 2014) proposes a macroscopic model to value users as a function of the advertising traffic their content generates. Here, in contrast, we wish to analyze social clicks at the level of a post, as we aim at understanding their propagation and interaction property, while previous analyses of social clicks centered on their distribution among URLs (Gabielkov et al. 2016).

A common motivation of our work and several others is to study propagations to quantify influence online (Cha et al. 2010; Bakshy et al. 2011), how news sharing is affected by social networks (An et al. 2011; 2014), and various mechanisms and drivers behind retweeting links (Kwak et al. 2010; Boyd, Golder, and Lotan 2010). Our work complements this line of work as it makes it possible to analyze reading habits, which was previously ignored. We also prove that it may reevaluate the strength of an influencer.

6 Conclusion

We reproduced and validated a publicly accessible method of analyzing social news consumption on Twitter. We developed a model for estimating CPI rates from publicly available data. We applied this new estimator to analyze the dynamics of a content publisher, `www.buzzfeed.com`. In our analyses, we revealed that there is a negative relationship between retweeting and click through rate. Furthermore, we revealed that beyond a threshold of mass sharing, the positive relationship between retweeting and absolute clicks dissipates. Further examination into the effects of sharing could provide greater insight into influence dynamics and influence maximization. In addition, future study into various types of content and sources could provide greater under-



(a) Retweets < 100: Strong positive correlation in both clicks and impressions

(b) Retweets > 100: Positive correlation in impressions, but no correlation in clicks.

Figure 7: Retweets vs Clicks*, Impressions*: The effect of sharing on clicks and impressions, at different sharing magnitudes.

standing into the general trend. Overall, we hope that our model and methodology will help foster better understanding of sharing and audience dynamics.

References

An, J.; Cha, M.; Gummadi, K.; and Crowcroft, J. 2011. Media landscape in Twitter: A world of new conventions and political diversity. In *Proceedings of the International Conference Weblogs and Social Media (ICWSM)*, 18–25.

An, J.; Quercia, D.; Cha, M.; Gummadi, K.; and Crowcroft, J. 2014. Sharing political news: the balancing act of intimacy and socialization in selective exposure. *EPJ Data Science*.

Bakshy, E.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Everyone's an influencer: quantifying influence on twitter. In *WSDM '11: Proceedings of the fourth ACM international conference on Web search and data mining*. ACM Request Permissions.

Boyd, D.; Golder, S.; and Lotan, G. 2010. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference*, volume 0, 1–10. Honolulu, HI: IEEE.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. 2010. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the International Conference Weblogs and Social Media (ICWSM)*.

Dave, V.; Guha, S.; and Zhang, Y. 2012. Measuring and fingerprinting click-spam in ad networks. In *SIGCOMM '12: Proceedings of the ACM SIGCOMM 2012 conference on Applications, technologies, architectures, and protocols for computer communication*, 175. New York, New York, USA: ACM Request Permissions.

Farahat, A., and Bailey, M. C. 2012. How effective is targeted advertising? In *WWW '12: Proceedings of the 21st international conference on World Wide Web*. ACM Request Permissions.

Gabielkov, M.; Ramachandran, A.; Legout, A.; and Chaintreau, A. 2016. Social Clicks: What and Who Gets Read on Twitter? In *ACM SIGMETRICS / IFIP Performance 2016*.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is Twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*. ACM.

McMahan, H. B.; Holt, G.; Sculley, D.; Young, M.; and Ebner, D. 2013. Ad Click Prediction: a View from the Trenches. *KDD '13: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Myers, S.; Zhu, C.; and Leskovec, J. 2012. Information Diffusion and External Influence in Networks. *KDD '12: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Saez-Trumper, D.; Liu, Y.; Baeza-Yates, R.; Krishnamurthy, B.; and Mislove, A. 2014. Beyond CPM and CPC: Determining the Value of Users on OSNs. *COSN '14: Proceedings of the 2nd ACM conference on Online social networks*.

Wong, D. 2015. Social media drove 31.24% of overall traffic to sites.