Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)

The Million Tweets Fallacy: Activity and Feedback Are Uncorrelated

Dan Vilenchik
Ben-Gurion University of the Negev
Israel

Abstract

In this paper we point to a surprising phenomenon in Online Social Networks (OSNs), which we call "The Million Tweets Fallacy". Our hypothesis is that the measurements of activity of a user in an OSN are not correlated with the measurements of feedback that the user receives on that activity. For example, the number of tweets is uncorrelated with the number of retweets or likes. In other words, a voluminously tweeting user is not necessarily fuelled by the attention he gets from his followers. An innovative aspect of this work is that we treat "activity" and "feedback" as multidimensional axes, and do not reduce the problem to a one-dimensional pairwise correlation problem. We apply our methodology to six OSNs. For Twitter, Instagram, LinkedIn and Steam we gathered the data ourselves, collecting features that cover both users' activity and feedback in the OSN. For YouTube and Flickr we used existing data from the literature. In all OSNs, with the only exception of Steam, we confirmed our hypothesis.

Introduction

The blood circulation in online social networking sites (OSNs) consists of three fundamental elements, *producing content*: posting opinions, questions, answers, photos, videos; *consuming content*: viewing videos, reading posts; *giving feedback*: liking, retweeting, sharing. Producing enables consuming, consuming leads to feedback, which in turn encourages producing. Thus, successful social networking sites manage to promote a healthy cycle of content creation and consumption. However, peculiar and non-intuitive phenomena may interfere with this cycle. For example the "Million Followers Fallacy", a term coined by Avnit (Avnit 2009), who pointed to anecdotal evidence that some users follow others simply because it is polite to follow someone who is following you. As a result the OSN contains supposedly-central users with a huge amount of followers that have limited interest in their posted content. This was confirmed in Twitter by showing that a user's number of followers and influence (measured as the ability to spread popular news topics) were not correlated (Cha et al. 2010). More generally, the claim that various simple statistics of the user's profile are not good indicators of influence and centrality were noted for example in (Trusov, Bodapati, and Bucklin 2010) or (Green October 6 2008).

In this work we extend these observations and point to a phenomenon which we call "The Million Tweets Fallacy". This term stands for the claim that measurements of a user's activity in the OSN are not correlated with the measurements of feedback that the user receives on that activity. More specifically, user statistics in an OSN may be viewed along two axes: the activity axis and the feedback axis. By activity we mean both statistics of content activity (e.g. total number of posts, number of posts per day, video vs. text, etc) and statistics of social activity (e.g. the number of friends, the number of users one follows, the number of likes that a user gives). The feedback axis consists of statistics that measure the feedback that a user receives on his activity, e.g. number of views, number of likes, number of retweets, number of users that follow that user, etc. It is important to note that our features do not concern the actual content that the user publishes, or the content of the feedback (when relevant). The following hypothesis formalizes the The Million Tweets Fallacy:

Hypothesis. *In an OSN, the scores of users in the activity axis are not correlated with the scores in the feedback axis.*

Figure 1 illustrates the claim in data that we collected for Twitter. The figure depicts two longs tails: very active users with very little feedback, and the other way around. Additionally, the massive bulk of users displays an incoherent mixture of activity-feedback ratios.

Coming up with a statistical framework that can confirm such an hypothesis may be a challenging task in itself, since every axis is in fact a multidimensional axis. One possibility is to reduce the multidimensional problem into a set of pairwise one-dimensional correlation tests. Doing so one naively assigns the same importance to all features, which might lead to erroneous or irrelevant conclusions. In addition, richer insights may be gained if axes remain multidimensional. Unfortunately, there is no straightforward way to compute a one-number correlation score for multidimensional random variables. Indeed a covariance matrix is the typical extension to the multidimensional setting.

We suggest a suitable statistical framework, based on Principal Component Analysis (PCA), that can confirm our hypothesis in a multidimensional fashion. The Principal Components (PCs) of the covariance matrix of a dataset, are carefully chosen linear transformations of the original set of features. The linear transformations give large weight to important features, and small weight otherwise (importance is measured, as customary, by variance). The PCs may be interpreted as a new set of complex features, forming new axes along which the data is redrawn. The complex features pertain to various users' behaviors and modes of interactions in the OSN. What makes PCA suitable for our goal is, first the fact that a PC is simply a vector, and hence from an algebraic point of view – a one dimensional object. However semantically it is a suitably chosen summary of a multidimensional random variable. Second, the values (scores) of the data points along different PC-axes are statistically uncorrelated (Lemma 1).

Using our statistical framework we checked the hypothesis in six different OSNs. For four OSNs, Twitter, Instagram, LinkedIn and Steam, we collected the data ourselves using standard crawling techniques (Mislove et al. 2007). We chose those networks because they represent a good variety: Twitter is a social text-based platform, while the content in Instagram is primarily visual and it is popular among a younger crowd. Steam and LinkedIn are thematic-niche OSNs; Steam is an online gaming community and LinkedIn has the job-market orientation. For YouTube and Flickr we use the data that was collected and analyzed in (Canali, Casolari, and Lancellotti 2012). For all networks, with the only exception of Steam, we confirmed our hypothesis. Let us note that our PCA-based statistical framework is sound but not complete, i.e. the hypothesis can be true for a certain OSN but still not confirmable using our framework. Steam may be one such example.

Related Work

Traditionally, information spreading in OSNs is viewed as flowing from key members to their followers, e.g. (Kozinets et al. 2010) or (Goldenberg et al. 2009). In contrast, the modern view of information flow emphasizes the importance of prevailing culture more than the role of influentials. Some researchers claim that people make choices based mainly on their peers' and friends' opinions rather than on influentials (Domingos and Richardson 2001). The dissonance between the two views was captured in popular science as the "Million Followers Fallacy" (Avnit 2009) and confirmed for Twitter (Cha et al. 2010). Concomitantly, some scholars explored alternatives to the structural centrality measures. Trusov *et al.* (2010) studied the connection between network activity of members and their influence in a certain OSN. They concluded that simple metrics, such as friend count and profile views, are likely to be inadequate proxies for user influence. Practitioners have also noticed the inability of simple metrics to capture influence on social network sites (Green October 6 2008). Our work may be viewed as a continuation of this line of research where feedback serves as a proxy for influence. We show that simple statistics of the activity plane are uncorrelated with feedback, and therefore might be bad predictors for feedback and in turn influence.

Another related research topic is the task of users characterization in OSNs. A PCA-based method was recently studied by Canali *et al.* (2012) with very promising results for YouTube and Flickr. Their main result is that the top PCs in both OSNs encode labels that correspond to measures of popularity and activity in the network. In this manner the PCs induce a soft classification of the users, in the sense that there is no single label per user but a continuum along each PC-axis. We extend this line of research to additional OSNs. Our feedback axis is the equivalent of the popularity axis of (Canali, Casolari, and Lancellotti 2012), and our activity axis (or axes, if several PCs are relevant) may be inspected to determine which specific type of activity is encoded by the PCs.

Methodology

In this section we present our statistical framework to verify the hypothesis *activity and feedback are not correlated*. The following lemma is a key observation. We use $\hat{\Sigma}$ for the (sample) covariance matrix, and X for the $n \times p$ data matrix (n is the number of samples and p the number of features).

Lemma 1 *Let* $\mathbf{v}_i, \mathbf{v}_j$ *be two PCs of* $\hat{\Sigma}$ *with* $i \neq j$*. The scores* $\mathbf{y}_i = X\mathbf{v}_i$ *and* $\mathbf{y}_j = X\mathbf{v}_j$ *satisfy* $\mathbf{y}_i^T \mathbf{y}_j = 0$*, i.e. they are uncorrelated.*

The proof, omitted here, follows immediately from definitions and the orthogonality of the PCs. Figure 1 illustrates what uncorrelation looks likes in our Twitter dataset.

Next we define two directions in the feature space that correspond to *activity* and *feedback*. Let $I_{act} \subset \{1, \ldots, p\}$ be the indices of the features that concern the activity of the user and similarly define I_{fdbk}. We assume that $I_{act} \cap I_{fdbk} = \emptyset$, i.e. every feature belongs at most one of the sets (some features may belong to neither e.g. *age*). For every PC $\mathbf{v}_i$ we define its "energy" in either directions (feedback and activity). Specifically, let $\mathbf{v}_i[j]$ be the j^{th} entry of $\mathbf{v}_i$. The energy is given by the sum of entries squared:

$$\alpha_i = \sum_{j \in I_{fdbk}} (\mathbf{v}_i[j])^2, \quad \beta_i = \sum_{j \in I_{act}} (\mathbf{v}_i[j])^2. \qquad (1)$$

The total energy of $\mathbf{v}_i$ is 1 as $\mathbf{v}_i$ is a unit vector, therefore $\alpha_i + \beta_i \leq 1$.

To see how these definitions are used to confirm the hypothesis we describe an idealistic scenario, which is too naive to be typical but very instructive. Suppose that $\mathbf{v}_1$ and $\mathbf{v}_2$ explain 100% of the variance. Namely, together they characterize 100% of the behaviour patterns (variance) of the social network users (according to the selected features). Further assume that $\alpha_1 = 1, \beta_1 = 0$ and $\beta_2 = 1, \alpha_2 = 0$. Namely, $\mathbf{v}_1$ points fully in the activity direction and $\mathbf{v}_2$ fully in the feedback direction. Lemma 1 implies that the behaviour in the $\mathbf{v}_1$-axis is uncorrelated with the behaviour in the $\mathbf{v}_2$-axis, and the hypothesis is confirmed.

Next we list a set of rules that quantify how close the data is to this idealistic scenario. We do so in two stages. First, identify which PCs are relevant (in the idealistic example, all but the first two PCs explained zero variance and are clearly irrelevant). Then, for relevant PCs, decide if their α_i and β_i values significantly classify them as *activity* or *feedback* (in

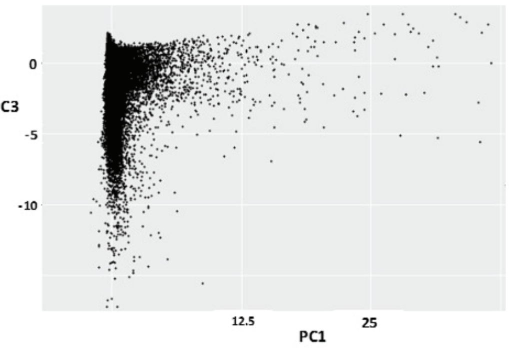

Figure 1: ≈285,00 Twitter users plotted according to their projections on PC1 (Feedback axis) and PC3 (Activity axis). This figure provides a clue of how uncorrelation looks like, or a visualization of Lemma 1.

the idealistic example this task was also trivial as α_i and β_i were binary).

To select relevant PCs we use the Guttman-Kaiser criterion (Yeomans and Golder 1982). Namely, a PC is considered relevant if it explains more than $1/p$-fraction of the variance, otherwise it is treated as explaining incidental variance or noise. To get a baseline for the values of α_i and β_i we study their distribution had the vector $\mathbf{v}_i$ been a random unit vector. Computing the expected values of α and β is easy: by symmetry the expected support of a random vector on a subset of indices I is simply $|I|/p$. We further compute, empirically, the $q = 0.025, 0.25, 0.75, 0.975$ percentiles and denote them by x_q^α and x_q^β. Note that by symmetry, in all these calculations only the sizes $|I_{fdbk}|$ and $|I_{act}|$ matter. In addition, different datasets may have different baselines since the ratio of $|I_{fdbk}|$ to $|I_{act}|$ may be different and the value of p may differ as well. We say that a PC $\mathbf{v}_i$ is:

- *Purely feedback* if $\alpha_i > x_{0.975}^\alpha$ and $\beta_i < x_{0.025}^\beta$
- *Purely activity* if $\beta_i > x_{0.975}^\beta$ and $\alpha_i < x_{0.025}^\alpha$
- *Neutral* if $\alpha_i \in [x_{0.25}^\alpha, x_{0.75}^\alpha]$ and $\beta_i \in [x_{0.25}^\beta, x_{0.75}^\beta]$
- Otherwise, we say that $\mathbf{v}_i$ is *mixed*.

Let $\mathcal{P} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r\}$ be the set of PCs which explain more than $1/p$-fraction of the variance. We use the following rules to decide the validity of the hypothesis:

1. If $\mathcal{P}$ contains only pure PCs and at least one of each type then we say that the hypothesis is confirmed.
2. Otherwise, if $\mathcal{P}$ contains a mixed PC or is missing at least one of the pure types, then we declare the framework unsuitable to validate the hypothesis.
3. Otherwise, if $\mathcal{P}$ contains neutral PCs, remove them from $\mathcal{P}$ and decide according to Rule 1.

The α, β-values of neutral vectors are typical of random vectors (using the inter-quartile measure). Hence they are dismissed in Rule 3 as explaining incidental variance with respect to activity or feedback patterns.

	PC1	PC2	PC3	PC4	PC5
Twitter	18%	16%	13%	10%	8%
Instagram	29%	19%	10%	9%	8%
LinkedIn	25%	11%	10%	6.5%	6%
Steam	27%	14%	10%	9%	8%

Table 1: The percentage of explained variance per PC. Last two columns correspond to variance below $1/p$. Accumulative variance of the top three PCs is between 46% and 57% in all four OSNs.

Data Collection

We crawled the network in a snowball approach, which is commonly used in the literature (Mislove et al. 2007). Crawling starts from a list of randomly selected users and proceeds in a BFS manner. At each step the crawler pops a user v from the queue, explores its outgoing links and adds them to the queue. In Twitter there is a link from v to w if v follows w. In Instagram the set of friends is private in most cases. We say that w is an outgoing link from v if w commented on v's pictures. In Steam the list of friends is public. In LinkedIn the list of friends (called connections) is private. As a proxy for v's friends we used the "People Also Viewed" box which tells what recent profiles w were viewed by people who viewed v.

We collected between 11 to 15 features per network, which were split roughly half-half between activity and feedback. Feedback features included for example the number of users following me, the number of retweets of my tweets by others, the number of likes I received or comments left on my pictures. The activity features included the volume of activity (e.g. posts per day, total number of posts), activity types (e.g. percentage of video vs pictures, urls vs. pure text), social activity (number of friends, number of likes I gave, number of tweets I retweeted). Similar features were used to find influential users in MySpace and Facebook (Eirinaki, Monga, and Sundaram 2012) or in YouTube and Flickr (Canali, Casolari, and Lancellotti 2012). We collected a total of 284,758 Twitter accounts, 52,574 in Instagram, 127,830 in Steam and 12,000 in LinkedIn. Different numbers stem from varying levels of technical difficulty in crawling each network and from time constraints. Twitter data is available at (Vilenchik and Yichye 2016).

Results

We start by fixing the relevant PCs. Table 1 shows the percentages of explained variance of the top PCs in all four OSNs. In all datasets, only the top three PCs pass the Guttman-Kaiser criterion, and hence considered for the sake of the hypothesis. Table 2 summarizes the statistics of the α_i and β_i values of the top three PCs across all OSNs. For Twitter we see that all vectors are pure: PC1 is purely feedback, PC2 and PC3 are purely activity. There are no mixed or neutral vectors, hence the hypothesis is readily confirmed. Similarly for Instagram, PC1 is purely feedback and PC2 and PC3 are purely activity. No neutral or mixed PCs, and again the hypothesis is readily confirmed. For LinkedIn, PC1 is a neutral vector, PC2 purely feedback and PC3 purely activity. Again we confirm the hypothesis. Note that the standard deviations of α and β are larger than in the other OSNs. This may be attributed to the fact that for LinkedIn we had a significantly smaller dataset. The situation for Steam is different than all the other networks: all three PCs are neutral. In other words, the directions that the top three PCs point to are not aligned with either Activity nor Feedback but rather a mixture of both. Hence we are in a position where we cannot confirm the hypothesis. Indeed, looking at the top users in PC1 reveals both heavy gamers that have a narrow social circle and low feedback (e.g. the user *marula79* who played 230 hours in the past 2 weeks, earned 4,000 badges, but has only 421 friends and received 164 comments on his profile) and light gamers that have a wide social circle and high feedback (e.g. the user *addmebby* who played merely 32 hours in the past 2 weeks, but has 1,677 friends and received 2,300 comments on his profile), and the spectrum in between.

Twitter	α	α avg	β	β avg
PC1	0.98	0.93 ± 0.09	0.02	0.07 ± 0.08
PC2	0.003	0.06 ± 0.08	0.997	0.94 ± 0.08
PC3	0.02	0.02 ± 0.01	0.98	0.98 ± 0.01
Percentiles	0.025	0.25	0.75	0.975
α	0.023	0.121	0.351	0.627
β	0.379	0.648	0.882	0.978

Instagram	α	α avg	β	β avg
PC1	0.996	0.99 ± 0.01	0.004	0.01 ± 0.03
PC2	2e-05	0.001 ± 0.002	0.99998	0.99 ± 0.003
PC3	0.003	0.01 ± 0.02	0.997	0.99 ± 0.02
Percentiles	0.025	0.25	0.75	0.975
α	0.055	0.213	0.490	0.751
β	0.235	0.506	0.783	0.943

We computed the α, β-values for two additional OSNs, Flickr and YouTube, using the detailed PCA results in (Canali, Casolari, and Lancellotti 2012). The YouTube data was collected in 2009 and contains nearly two million users. The top three PCs explain 60% of the variance. The Flickr data was collected in 2011. The top four PCs explain 78% of the variance. We found that in YouTube, PC1 is purely feedback, PC2 purely activity and PC3 is neutral. In Flickr, PC1 is neutral, PC2 is purely feedback and PC3,PC4 are purely activity, and again the hypothesis is confirmed in both OSNs.

Acknowledgements

This work was supported by the ISF grant number 1388/16. We thank Maor Abutbul, Barak Yichye, Shalom Toledo and Yaniv Costica for helping with the data collection.

LinkedIn	α	α avg	β	β avg
PC1	0.27	0.36 ± 0.16	0.73	0.64 ± 0.17
PC2	0.62	0.42 ± 0.23	0.38	0.57 ± 0.23
PC3	0.03	0.09 ± 0.09	0.97	0.9 ± 0.09
Percentiles	0.025	0.25	0.75	0.975
α	0.0406	0.149	0.361	0.615
β	0.387	0.638	0.852	0.958

Steam	α	α avg	β	β avg
PC1	0.16	0.18 ± 0.03	0.84	0.81 ± 0.03
PC2	0.1	0.11 ± 0.03	0.9	0.88 ± 0.03
PC3	0.2	0.15 ± 0.06	0.8	0.85 ± 0.06
Percentiles	0.025	0.25	0.75	0.975
α	0.005	0.062	0.264	0.550
β	0.442	0.734	0.938	0.994

Table 2: In each table the top part is the values of α and β computed via Eq.(1). The average is taken over 100 random subsamples each of size 5,000-10,000 users (depending on the OSN). The bottom part are empirically computed percentiles over a sample of $10,000$ random unit vectors.

References

Avnit, A. 2009. The million followers fallacy. *http://blog.pravdam.com/the-million-followers-fallacy-guest-post-by-adi-avnit*.

Canali, C.; Casolari, S.; and Lancellotti, R. 2012. A quantitative methodology based on component analysis to identify key users in social networks. *Int. J. Social Network Mining* 1(1):27–50.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. P. 2010. Measuring user influence in twitter: The million follower fallacy. In *Proc. of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, 10–17.

Domingos, P., and Richardson, M. 2001. Mining the network value of customers. In *Proc. of the seventh ACM SIGKDD*, 57–66.

Eirinaki, M.; Monga, S. P. S.; and Sundaram, S. 2012. Identification of influential social networkers. *Int. J. Web Based Communities* 8(2):136–158.

Goldenberg, J.; Han, S.; Lehmann, D.; and Weon-Hong, J. 2009. The role of hubs in the adoption process. *J. of Marketing* 73(2):1–13.

Green, H. October 6, 2008. Google: Harnessing the power of cliques. *BusinessWeek* 50.

Kozinets, R.; de Valck, K.; Wojnicki, A.; and Wilner, S. 2010. Networked narratives: Understanding word-of-mouth marketing in online communities. *J. of Marketing* 74(2):71–89.

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

Trusov, M.; Bodapati, A.; and Bucklin, R. 2010. Determining influential users in internet social networks. *J. of Marketing Research* 47(4):643–658.

Vilenchik, D., and Yichye, B. 2016. Twitter data. *https://github.com/sdannyvi/TwitterDataAnon*.

Yeomans, K., and Golder, P. 1982. The guttman-kaiser criterion as a predictor of the number of common factors. *The Statistician: J. of the Institute of Statisticians* 31:221–229.