CO-438: Term Paper Presentation

•••

Topic: “Simple statistics are sometime too simple: A case study in social media data” (Issue 2, Feb 2020)

By Syed Husain Mustafa
Faculty: 16PEB276
Enrol: GJ6496

Index

1. Abstract

Simple statistics are often insufficient at explaining a users' behavior on Social Media Platforms. As such relying purely on simple bookkeeping statistics such as the number of likes, retweets, etc does not explain a users' contribution to the online social circle, and can lead to a selective inference problem. To alleviate these issues the author (D. Vilenchik) proposed a novel statistical framework dubbed as *Semantic Shattering* that can detect semantic inconsistencies in the data.

The framework has been applied to measure the relation between User Activity and Community Feedback from 6 platforms, and leads to a surprising counter-intuitive finding in 3 of them, ie; Youtube, Instagram, and Twitter.

2. Introduction

- Making sense of social media is challenging not only due to the shear volume of data but also because simple statistical data does not lead to useful conclusions.
- In a study conducted by Rao et. al. targeted at predicting latent user attributes in Twitter such as gender, age, political orientation, and regional origin, simple statistics were found to be of no use.
- Cha et. al. also studied Twitter and found that users who have many followers, those typically thought of as popular users, are not necessarily influential in terms of spawning retweets or mentions. There is only a 12% overlap between the top 90% of most followed users, and the top 90% of most retweeted users.

Million Follower Fallacy

An explanation to the curious case of popular users rarely being influential was given by Avnit.

The research points to anecdotal evidence that some users follow others because it is polite to follow back. However, this latent feature of politeness only gives a partial answer.

If we choose to measure popularity using higher dimensions it is uncertain if the fallacy will persist.

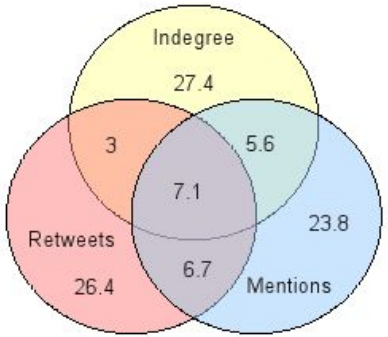

Figure 1: Venn diagram of the top-100 influentials across measures: The chart is normalized so that the total is 100%.

Image Source: Perez, Sarah "The Million Follower Fallacy: Audience Size Doesn't Prove Influence on Twitter" *readwrite.com*. Web Technology Blog, 19 March 2010.

2a. The Authors' Contribution

- *Semantic Shattering* is a multidimensional framework to detect semantic inconsistencies in the data, such as the one pointed by the Million Followers Fallacy.
- More formally, for data collected over m features $f_1, \ldots, f_m$ *if there exist* r qualities of interest $Q_1, \ldots, Q_r$ defined using these features, then we say the dataset D semantically shatters if there exist two qualities Q_i and Q_j which are not correlated.
- In the case of the Million Followers' Fallacy we can consider f_1 as the number of followers, f_2 as the number of retweets, Q_1 as the quality of Popularity measured by f_1 and Q_2 as Influence measured using f_2 .

- While there are clearly qualities that naturally do not correlate, such as age and the average number of a's in ones post. *Semantic Shattering* is interesting when qualities that one naturally assumes to be correlated - shatter, such as Popularity and Influence, or as found in this research: Activity and Feedback.
- Developing a statistical framework to check for *semantic shattering* is a non-trivial task since every quality may be a multidimensional random variable and correlation is not defined in this case.
- This problem of multidimensionality can be overcome by limiting the qualities to a single feature, however, this limits the descriptive power of the qualities, and raises validity issues regarding the conclusions.
- Another approach is to use pairwise 1-dimensional correlation tests. This leads to all features having the same level of importance, which results in erroneous conclusions. This may also cause selective inference problems.

Selective Inference Problem

Consider if we are required to judge the strength of a relation between “A” and “B”. Where “A” and “B” are two features. If we are provided with the information that “A” and “B” have a correlation of 0.9 - that is probably noteworthy. However, if the value of 0.9 was arrived at by comparing correlation between pairs of “A” and “B” in a set of 1000 entries and found that the single highest value was 0.9 - the value does not hold much weight.

Hence, if we’re told simply that corr(A,B)= 0.9 without knowing how the value was obtained we do not have enough insight to give that value any significance. This statistical problem is known as the *selective inference problem.*

- The *Semantic Shattering* framework is based on Principal Component Analysis that checks for semantic shattering in a multidimensional fashion.
- The usefulness of the framework is demonstrated by revealing a new kind of shattering in Social Networks between Q1 = Feedback and Q2 = Activity
- Both qualities are captured using simple statistics such as total number of posts, number of posts per day, the number of friends, the number of users one follows, the number of likes that a user gives. Feedback includes number of views, number of likes, retweets, number of followers, etc.
- Data from 4 social networks, ie; Twitter, Instagram, LinkedIn and Steam was collected using a snowball crawling approach. While data from prior studies was used for Flickr and Youtube.

- It was observed that the Principal Components often point in directions that correspond to characteristics of users, say a measure of popularity or influence.
- Paradoxically, whenever the PCs point in a “useful” direction this gives rise to semantic shattering. This property is inherent to the PCA method, as projections along different Principal Components are uncorrelated(per Lemma 1).
- The Activity-Feedback Shattering was observed with Youtube data, where users were not adequately compensated by their social circle for positive activities, ie; viewing videos, liking videos, etc. Similar trends were seen with Instagram and Twitter.
- The author asserts that these generic/ multipurpose social networks do not cater to a specific niche of content, hence their audience is not as invested or interested in recipricating their efforts in kind.

- Social Media Platforms like Steam, Flickr, and LinkedIn however did not display a similar trend, ie; there was no semantic shattering. The author asserts that since these platforms cater to a more thematic-niche and the content shared by a user reflects activities that require a considerably larger effort: gaming, semi-professional photography, and one's career and education, the social circle is more likely to provide good feedback for good user activities.
- Indeed, sociologists typically use the concept of commitment when they are trying to account for the fact that people engage in a consistent manner.

3. Methodology

The Principal Components (PCs) of the covariance matrix are carefully chosen linear transformations of the original set of features. As such, the PCs may be interpreted as a new set of complex features, forming new axes along which the data can be redrawn. As per the authors' terminology, the PCs represent the qualities with respect to which semantic shattering may occur.

3a. Deriving Semantic Shattering

The first key observation is summarized in the following lemma. We use X for the $n \times p$ data matrix (n is the number of samples and p the number of features) and $\Sigma = \frac{1}{n} X^T X$ for the sample covariance matrix. Bold font is used for vectors, and they are considered as column vectors.

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

3.1.0.1 Proof: The proof follows immediately from definitions.

$$\mathbf{y}^T_i \mathbf{y}_j = (X\mathbf{v}_i)^T(X\mathbf{v}_j) = (\mathbf{v}^T_i X^T)(X\mathbf{v}_j) = \mathbf{v}^T_i (X^T X)\mathbf{v}_j = n\mathbf{v}^T_i \Sigma \mathbf{v}_j.$$

Since $\mathbf{v}_j$ is an eigenvector of Σ we can substitute $\Sigma\mathbf{v}_j$ with $\lambda_j\mathbf{v}_j$ and obtain

$$n\mathbf{v}^T_i \Sigma \mathbf{v}_j = n\mathbf{v}^T_i \lambda_j \mathbf{v}_j = (n\lambda_j)\mathbf{v}^T_i \mathbf{v}_j = 0 \quad (1)$$

The last equality is due to the orthonormality of the eigenvectors. Figure 1 illustrates Lemma 1 in the Twitter dataset.

Theorem 2 sets up the formal framework to measure correlation between qualities. It contains three easily computable sufficient conditions for a dataset ***D*** to exhibit semantic shattering. For a vector $\mathbf{v}$, we use the notation *supp(**v**)* for the support of $\mathbf{v}$, namely *supp*($\mathbf{v}$) = {r : $\mathbf{v}$[r] 6= 0}.
Abusing notation, we use Q_i also for the set of features that define this quality.

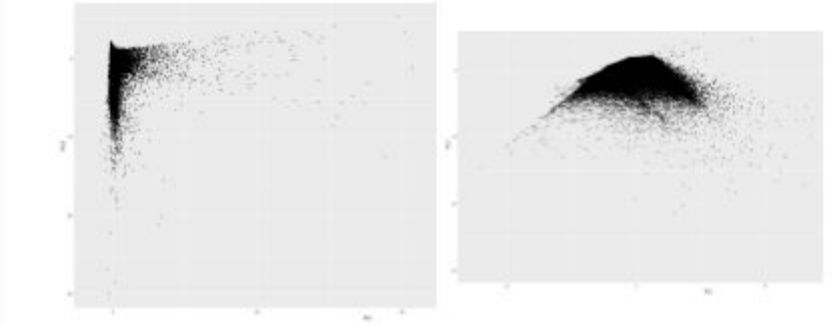

Fig. 1: A projection of 285,000 Twitter users on two PCs: left pane it is PC1 (Feedback) and PC2 (Activity) and right pane it is PC2 and PC3 (also Activity).

Theorem 2. Let $\mathbf{v}_1, \ldots, \mathbf{v}_p$ be the PCs of the covariance matrix of a p-dimensional dataset D. Let $Q_1, \ldots, Q_r$ be the qualities spanned by the p features. If there exist two qualities Q_s, Q_t that satisfy the following conditions, then ***D*** exhibits semantic shattering.

1) $Q_s \cap Q_t = \varnothing$.
2) Let $A = \{\mathbf{v}_{i1}, \ldots, \mathbf{v}_{ia}\}$ be the set of PCs that satisfy $supp(\mathbf{v}) \subseteq Q_s$ and let $B = \{\mathbf{v}_{i1}, \ldots, \mathbf{v}_{ib}\}$ be those that satisfy $supp(\mathbf{v}) \subseteq Q_t$. Then, $A \neq \varnothing$ and $B \neq \varnothing$.
3) For every $k = 1, \ldots, p$, either $\mathbf{v}_k \in A \cup B$ or $supp(\mathbf{v}_k) \cap (Q_s \cup Q_t) = \varnothing$.

2.1.0.2 Proof:: The third condition implies that all the information in the dataset regarding qualities Q_s, Q_t is in the vector space spanned by $A \cup B$. Therefore vectors $\mathbf{v}_k \notin A \cup B$ may be ignored. The first and second conditions imply that the vector space spanned by A contains all the information about Q_s and similarly the vector space spanned by B contains Q_t. ***Lemma 1*** applied to all the pairs $\mathbf{v}_i \in A$ and $\mathbf{v}_j \in B$ characterizes the manner in which Q_s and Q_t are uncorrelated.

Theorem 2 implicitly assumes that the qualities are indeed defined by the corresponding features, and that every feature is relevant to at most one quality.

3b. Relaxing Theorem 2

The setting of Theorem 2 is too "clean" to be relevant for real data. One major obstacle is that every PC $\mathbf{v}$ will typically satisfy $supp(\mathbf{v}) = \{1, \ldots, p\}$, yet because numerical reasons, making it impossible to meet the condition of the theorem. Therefore in order for the framework to be useful it needs to be relaxed. The relaxation that proposed is the following. For simplicity of presentation we assume that there are just two qualities Q_1, Q_2 and that the features split among them. The extension to the general case is straightforward. For every PC $\mathbf{v_i}$ we define its' *energy* in the direction of Q_1, Q_2 as follows:

$$\alpha_i = \sum_{r \epsilon Q_1} (\mathbf{v}_i[r])^2 \quad \beta_i = \sum_{r \epsilon Q_2} (\mathbf{v}_i[r])^2$$

The total energy of every $\mathbf{v_i}$ is 1 as it is a unit vector, therefore $\alpha_i + \beta_i = 1$. The conditions of Theorem 2 are equivalent in this case to requiring that for every PC $\mathbf{v_i}$, either $\alpha_i = 1$ or $\beta_i = 1$. We relax this requirement, by demanding that $\alpha_i \geq x_{0.975}^{\alpha}$ and $\beta_i \leq x_{0.025}^{\beta}$ or vice versa, where $x_{0.975}^{\alpha}$ is the 0.975^{th} percentile of the α value had the vector $\mathbf{v_i}$ been a random p-dimensional vector on the unit sphere. In the same manner all the other percentiles are defined. We call this the (α,β)- separation property. Note that the choice of 0.975 and 0.025 is somewhat arbitrary, and in general the closer the percentiles are to 1 or 0 the closer we are to the setting of Theorem 2.

The second issue is PCs that explain incidental variance or noise. Ignoring such PCs is a common Practice, known as the **Guttman-Kaiser (GK) criterion**, where a PC is considered informative only if it explains more than 1/p-fraction of the variance. However even among the PCs that pass the GK-criterion, some may fail to satisfy the (α,β)-separation property. If they are border-line with respect to GK-criterion and their α and β values fall in the inter-quartile regime, we classify them as Neutral, since intuitively their energy is spread between Q_1 and Q_2 as one would expect from a random vector. To summarize the PC classification, we say that a PC $\mathbf{v_i}$ is

- Purely Q1 if $\alpha_i > x^{\alpha}_{0.975}$ and $\beta i < x^{\beta}_{0.025}$
- Purely Q2 if $\beta_i > x^{\beta}_{0.975}$ and $\alpha i < x^{\alpha}_{0.025}$
- Neutral if $\alpha_i \in [x^{\alpha}_{0.25}, x^{\alpha}_{0.75}]$ and $\beta_i \in [x^{\beta}_{0.25}, x^{\beta}_{0.75}]$ and $\mathbf{v_i}$ explains roughly 1/p-fraction of the variance.
- Mixed in all other cases.

We say that a dataset $\boldsymbol{D}$ exhibit semantic shattering if:

- All the PCs of its covariance matrix that pass the GK-criterion are either purely Q_1, purely Q_2, or Neutral.
- There exists at least one pure PC for every quality. Otherwise, i.e. if there exists a mixed PC or at least one of the pure types is missing, then we declare that the framework was unable to detect semantic shattering.

4. Shattering in Online Social Media

4a. Data Collection

The network is crawled in a snowball approach, which is commonly used in the literature. 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 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 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.

Data from 284,758 Twitter accounts, 52,574 from Instagram, 127,830 from Steam, and 12,000 from LinkedIn. Different numbers stem from varying levels of technical difficulty in crawling each network and from time constraint. The Youtube data was collected in 2009 and contains roughly 1.7 million users. The Flickr data was collected in 2011 and contains roughly a million users.

4b. Results

For each of the 4 datasets the covariance matrix was computed, and PCA was performed. For each PC the percentage of variance captured was recorded. For Youtube and Flickr these numbers were taken from prior studies. Table-1 details the percentage of variance explained by every PC, from which we conclude which PCs pass the GK-criterion. There PCs were classified as pure Activity, pure Feedback, Neutral or Mixed according to the rules from Section 3.

	PC1	PC2	PC3	PC4	PC5	p
Twitter	18.15%	16.2%	13%	10%	(8%)	12
Instagram	29%	19.1%	10.15%	(9%)	(8%)	11
LinkedIn	25.1%	11.7%	10.3%	7.1%	6.8%	15
Steam	27%	13.6%	10.5%	9.7%	(8%)	11
YouTube	29%	19%	12%	(8%)	(7%)	11
Flickr	26.4%	19.7%	14.3%	12.1%	(8.7%)	9

TABLE 1: The percentage of explained variance per PC. Parenthesized numbers correspond to variance below $1/p$. The last column gives the number of features, p.

4b1. Twitter, Instagram and Youtube

Table 2 summarizes the statistics of the α and β values of the PCs that passed the GK-criterion. Using the classification rules we conclude that in Twitter PC1 is purely Feedback, PC2 and PC3 are purely Activity. There are no Mixed or Neutral PCs. Hence we confirm semantic shattering. Similarly in Instagram, PC1 is purely Feedback and PC2 and PC3 are purely Activity. No Neutral or Mixed PCs and again semantic shattering. In Youtube PC1 is purely Feedback and PC2 purely Activity. PC3 explains 12% of the variance, slightly more than 1/p= 1/11 ≃ 9%, and both its α and β-values fall in the inter-quartile regime. Therefore we classify it as Neutral, and confirm shattering.

Twitter	α	α avg	β	β avg
PC1	0.98	0.93 ± 0.09	0.02	0.07 ± 0.08
PC2	0.013	0.014 ± 0.08	0.997	0.96 ± 0.08
PC3	0.02	0.02 ± 0.01	0.98	0.98 ± 0.01
PC4	0.006	0.006 ± 0.008	0.994	0.99 ± 0.02
Percentiles	0.025	0.25	0.75	0.975
α	0.022	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
YouTube	α	α avg	β	β avg
PC1	0.92		0.08	
PC2	0.003		0.997	
PC3	0.35		0.65	
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

TABLE 2: The values of α (Feedback) and β (Activity) were computed for each PC using Eq.(2). The average is taken over 100 random subsamples each of size 5,000-10,000 users (depending on the social network). The percentiles were empirically computed over a sample of $10,000$ random unit vectors.

4b2. Linkedin, Steam and Flickr

Looking at Table 3 we see that in Linkedin, Steam and Flickr the leading PC is Mixed, namely neither pure Activity nor pure Feedback. This is a dead-end in terms of the semantic shattering framework hence shattering cannot be confirmed.
Nevertheless, we see that in LinkedIn PC2 is purely Feedback and PC3, PC4, PC5 are purely Activity. In Flickr PC2 is purely Feedback and PC3, PC4 are purely Activity. On the other hand, in Steam all four PCs are Mixed.

LinkedIn	α	α avg	β	β avg
PC1	0.32	0.36 ± 0.16	0.68	0.64 ± 0.17
PC2	0.6	0.42 ± 0.23	0.4	0.57 ± 0.23
PC3	0.09	0.11 ± 0.09	0.91	0.92 ± 0.08
PC4	0.08	0.08 ± 0.07	0.92	0.9 ± 0.1
PC5	0.08	0.09 ± 0.08	0.92	0.91 ± 0.09
Percentiles	0.05	0.25	0.75	0.94
α	0.092	0.2	0.44	0.6
β	0.387	0.638	0.852	0.897
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
PC4	0.11	0.13 ± 0.05	0.89	0.88 ± 0.04
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
Flickr	α	α avg	β	β avg
PC1	0.58		0.42	
PC2	0.999		0.001	
PC3	0.09		0.91	
PC4	0.025		0.975	
Percentiles	0.05	0.25	0.75	0.95
α	0.11	0.28	0.60	0.81
β	0.19	0.40	0.72	0.89

5. Semantic Shattering and User Classification

- Canali et al. used PCA to characterise users in Youtube and Flickr. Their main result is that the top PCs in both networks encode labels that correspond to measures of popularity and activity in the network. In this way 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.
- Viswanath et al. used PCA to classify Facebook users as either “normal” or “anomalous”.
- If we identify qualities with PC labels, then semantic shattering implies a PCA-based classification scheme. Indeed the top ten users in the Instagram dataset in PC1-measure include leading celebrities. PC2 and PC3 are pure Activity in Instagram and describe two different aspects of activity. Top users in PC2-measure provide specialized content such as fitness, photography, etc. Top users in PC3 measure include users that focus on video content: fashion and celebrity news, Islamic religious content and videos similar to America’s Funniest Home Videos TV show.
- Another example is PC2 in Twitter which was classified as purely Activity. The negatively signed features in PC2 are indicators of robot/spam behavior: *NumOfUrl* and *NumOfHashtag*. Indeed, the main way of spamming in Twitter is by hashtags and URLs, which appear in shortened form.

- PC3 in Twitter is also purely Activity but its support is dominated by other activity features. The main features of PC3 include the number of tweets and tweet rate, likes given to others, and the number of other users mentioning. These attributes measure the extent to which a user is a content provider. In addition the feature of retweets from other users appears in an opposite sign to the former, which excludes content providers that don't generate content but just share it. The top accounts in PC3-measure in our sample include news providers, and video game support.
- Steam is an example where all four PCs were Mixed, and in particular semantic shattering could not be verified. Indeed the Steam PCs were not useful for the classification task. For example, looking at the top users in PC1 reveals both heavy gamers that have a narrow social circle and low feedback and light gamers that have a wide social circle and high feedback, and the spectrum in between.

6. Discussion

- Given a dataset recording human activity it is often unclear how to go about identifying interesting phenomena in it. Lack of clear guidelines for analytic practice makes quantitative exploration of large-scale behavioral data more of an art than a science. This is a long-standing observation, which manifests in many settings, most notably Simpson's paradox, which is typically addressed in the context of supervised learning, and more specifically regression-type problems.
- The author has proposed an unsupervised learning framework to detect semantic inconsistencies that a dataset may possess due to unaccounted for confounding variables. This framework generalizes previous work, such as the Million Followers Fallacy.

- With this research the author concludes that simple statistics are useful in making sense of social media data to the extent that they are enough to classify users, but they are not always sufficient to obtain a good understanding of the nature of interaction between the different qualities.
- The author lists the following as the limitations of his method:
    - Neutral PCs are treated as if they did not met the Guttman-Kaiser criterion. This however, is a defendable choice, but it might be the case that a Neutral PC does encode a certain property that involves both Activity and Feedback. In this case ignoring a Neutral PC is wrong, and one needs to check which property it is and proceed accordingly.
    - The collected data does not cover all possible simple statistics of a user's profile, indeed no dataset can be that comprehensive. Therefore semantic shattering is restricted to the features that one selects. When friendship is a symmetric relationship, like in Facebook and Steam,it is not clear how to classify the features "number of friends". We chose to classify it as Feedback since friendship requires authorization and therefore provides feedback on the user.
- In light of these limitations, the author has considered to leave as an open thread the further refinement of this technique for future research.

7. References

[1] D.Vilenchik, “Simple statistics are sometimes too simple: A case study in social media data”, IEEE Transactions on Knowledge and Data Engineering, pp. 402-408, vol. 32, issue 2, Feb 1 2020.

[2]Taylor, J., and Tibshirani, R. J. (2015), “Statistical Learning and Selective Inference,” *Proceedings of the National Academy of Sciences*, 112, 7629–7634

[3] Rao D, Yarowsky D, Shreevats A, Gupta M. “Classifying Latent User Attributes in Twitter.” In: Proc of the 2nd Int. Workshop on Search and Mining User-generated Contents; 2010. p. 37–44.

[4] Cha M, Haddadi H, Benevenuto F, Gummadi KP. Measuring user influence in Twitter: The million follower fallacy. In: Proc. of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM); 2010. p. 10–17