Measuring Influence on Instagram: A Network-Oblivious Approach

Noam Segev
Klear
noam.segev@klear.com

Noam Avigdor
Klear
n@klear.com

Eytan Avigdor
Klear
e@klear.com

ABSTRACT

This paper focuses on the problem of scoring and ranking influential users of Instagram, a visual content sharing online social network (OSN). Instagram is the second largest OSN in the world with 700 million active Instagram accounts, 32% of all worldwide Internet users[1]. Among the millions of users, photos shared by more influential users are viewed by more users than posts shared by less influential counterparts. This raises the question of how to identify those influential Instagram users.

In our work, we present and discuss the lack of relevant tools and insufficient metrics for influence measurement, focusing on a network oblivious approach and show that the graph-based approach used in other OSNs is a poor fit for Instagram. In our study, we consider user statistics, some of which are more intuitive than others, and several regression models to measure users' influence.

CCS CONCEPTS

• **Information systems** → **Social recommendation**; *Social advertising*;

KEYWORDS

Social Media, Instagram, Influence, Ranking

ACM Reference Format:
Noam Segev, Noam Avigdor, and Eytan Avigdor. 2018. Measuring Influence on Instagram: A Network-Oblivious Approach. In *Proceedings of The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, July 8–12, 2018 (SIGIR '18),* 4 pages.
https://doi.org/10.1145/3209978.3210134

1 INTRODUCTION

The transition to Web 2.0 transformed the business models of online marketing from a global ad approach based to individual opinions and targeted campaigns [2, 23, 35, 40]. Web 2.0 not only took traditional marketing strategies to the extreme via viral marketing campaigns [31, 36, 43], but it also gave rise to new techniques of brand building and audience targeting via influencer marketing [12, 44]. In fact, the use of micro-influencers, trusted individuals within their communities, has been seen as a more effective way to build a brand in terms of audience reception and return on investment [9, 25, 32].

Instagram, which is a visual content sharing online social network (OSN), has become a focal point for influencer marketing. With power users and micro-influencers publishing sponsored content companies need to rate these influencers and determine their value[17, 18, 38]. Most of today's scoring themes rely on graph-based algorithms of a known network graph. Such graphs are not always available, and building them for Instagram users requires a great deal of resources, e.g., crawling time and computing costs. A possible solution would be to infer the underlying network structure using the user activity logs, as described by Barbieri et al.[7], but even in the event a graph is constructed it would not necessarily be of much use given that information decays exponentially along the graph even under optimal passive information propagation, which is not the case.

The rest of the paper is organized as follows: In Section 2 we described OSNs in greater detail as well as current influence measuring schemes. We then present our annotations and formal description of the problem of measuring and ranking influence in Section 3. The dataset of Instagram users and their posts is described in Section 4, followed by discussion on the extracted and aggregated features of the testable data in Section 4.2. Following this, we present our testing methodology, baselines, regression models and experimental results in Section 6. Finally, we discuss our conclusion and possible future work in Section 7.

2 BACKGROUND AND RELATED WORK

Online social media networks are often described as a directed graph with entities such as users acting as nodes and relationships as the edges. Such edges can be unidirectional or bi-directional, e.g., an Instagram "follower" and a Facebook "friendship", respectively. These edges do not need to represent a long-lasting relationship; they can signal a one-time engagement, e.g., a "like" or a "comment". Following this, link prediction in OSNs became an active research field focused on community detection, in the case of users as nodes [13, 27], or content suggestion otherwise [3, 37, 39].

In most OSNs, user-generated content is "pushed", i.e., propagated via interaction. When a user uploads a post, their followers can see the post and choose to pass it along, creating a pyramid-formed cascade of information. Thus, if user A follows user B who, in turn, follows user C, and user C posts some content user B chooses to share, user A is passively influenced by user C. These social micro-networks tend to grow around influential, active users [20, 33]. Instagram content, however, is "pulled", i.e., information propagation requires activity along the pyramid, such that, using

[1] https://www.omnicoreagency.com/instagram-statistics

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210134

our earlier example, for user C's post to reach user A, user A must look for content suggested by trusted users.

This situation raises the question of how to rank users in OSNs. As OSNs are traditionally described as graphs, ranking has been done using various graph statistics, from simple in/out degree to node closeness [15, 16], as is the case with the work of Anger and Kittl in Twitter [4] and Agarwal et al. in the context of influential blogs [1]. Other techniques extend to existing link analysis algorithms - the most popular one being PageRank [8, 11]. Weng et al. suggested twitterRank [42] and Khrabrov et al. introduced starRank [30], both extensions of PageRank working on Twitter's follower and engagements graphs, respectively. On LinkedIn, a professional OSN, Budalakoti and Bekkerman suggested a fair-bets model for ranking via transfer of authority [14], and on Instagram Egger suggested a PageRank extension for influencer ranking [19].

3 PROBLEM FORMULATION

The influence of a user in an OSN has been described either in simple, intuitive measures or as a non-intuitive measurable graph statistic with no real-world meaning [19, 24, 42]. One such measure is the user's expected post engagements. We extend this definition in the realization that being exposed to specific content often does not lead to active engagement.

We say the influence of an Instagram user (*Instagrammer*) is the expected exposure their content would receives, or, their expected number of views per post. Adhering to the law of large numbers, we can estimate the users' influence using Definition 3.1:

Definition 3.1. Let U be the set of all Instagrammers, C be all content posted on Instagram, v_c the number of Instagrammers that saw post $c \in C$ and $C_u \subset C$ is the content posted by $u \in U$. We say that the *influence* of Instagrammer u is:

$$Inf_u = \frac{\sum_{c \in C_u} v_c}{|C_u|}.$$

4 INSTAGRAM DATASET

For the purpose of this study, a set of Instagram data was prepared in April 2017, including posts published during 2015-2016 but prior to September 2016. We focused on a subset of Instagram posts where view counts were accessible. Independent studies have shown that 50% of engagements of an Instagram post happen within 72 minutes of publication and 95% within the following week[2]. As the change of feed ranking in March 2016 did not cause statistically significant changes to activity, and as all posts examined by us were over 6 months old, we say that the data is stable, meaning, all posts have reached at least 95% of their potential views and engagements. The data was prepared as follows:

(1) We gathered information on videos [3] published by a set of randomly selected Instagrammers with publicly accessible profiles. Denote the set of users as U. Each of these Instagrammers must have published a minimum of 10 video posts before September 2016.
(2) For each video $c \in C$, we collected the following metrics:

$likes_c$ Number of likes awarded to post c.
$comments_c$ Number of comments given to post c.
v_c Number of Instagrammers who watched part of the video.

A total of 940, 439 posts by 115, 044 Instagrammers was collected[4].

4.1 Instagram Statistics

The distribution for log average views per Instagrammer is presented in Figure 1a, from which we can tell that this statistic behaves in a log-normal distribution with a mean of 748 views. Furthermore, as this distribution is so close to normal, we ascertain that our selection of sampled Instagrammers is a good semblance of real-world influence with micro-influencers populating the dense mean and casual users and celebrities appearing at the distribution extremes.

Post views per followers and per engagement appear in Figures 1b and 1c, respectively; these show some underline truths of Instagram. It can be seen that normally, the number of followers a user has outnumber his views, as we expect following the described flow of information. However, we found that this is not the case for sponsored posts, massively engaged content or externally referenced content. Another unlikely situation is of posts having more engagements than views. This relates either to bought engagements, often via automation tools and fake accounts, or to an interesting phenomenon on Instagram known as "Like You, Like Me" where content is engaged simply to reciprocate prior engagements. The issue mitigates as the number of engagements increase.

To avoid these sorts of odd behaviors, we performed univariate outliers removal, ignoring the top and bottom posts for users with posts statistics above 2 standard deviations.

4.2 Features Collected

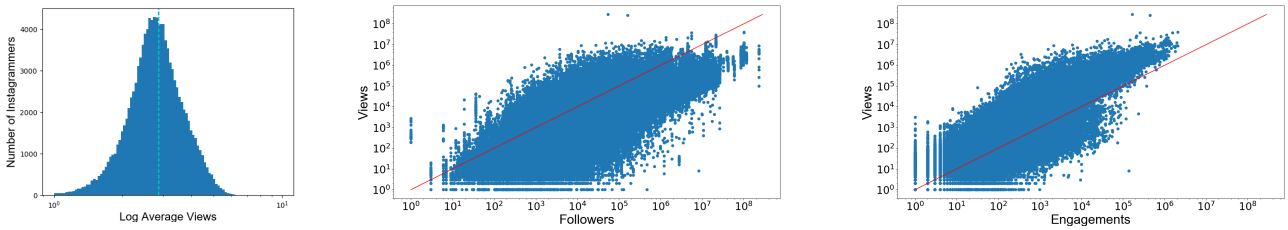
For the purpose of this work, we collected basic features directly from Instagram. Expanding on the posts features mentioned above, we also collected user specific statistics. We then considered each user as a data point with the following statistics:

$likes$ The average number of user post likes.
$comments$ The average number of comments per user post.
$followers$ The users audience size.
$\sqrt{likes \cdot followers}$ Geometric mean of likes and followers, taken as neither statistic is an exact representation of influence.
$\frac{followers}{post}$ Used to suggest odd behavior as same level influencers should have similar ratios.
$\frac{comments}{likes}$ Another odd behavior indicator as bought engagements tend to effect likes more than comments.
$focus$ The difference and ratio between most and least engaged post, these features were designed to test the variance and stability of a user engagement level.

5 REGRESSION MODELS

We attempt to measure influence using well known regression models via the features described at Section 4.2. Furthermore, as some models are sensitive to redundant features, we perform recursive feature elimination, generating a subset of informative features for the problem at hand.

[2] https://blog.takumi.com/the-half-life-of-instagram-posts-3db61fb1db75
[3] We used Instagram API to collect user statistics. We did not use the API to gather data for the posts themselves due to API limits. Instead, we parsed each post web-page.
[4] This collection of anonymized public information is available at https://klear.com/sigir/instagram_data.zip



(a) Views Histogram (b) Views per Followers (c) Views per Likes

Figure 1: Distributions per Instagrammer

The models tested include:

Ridge Regression(RR) An extension of Linear Regression, RR attempts to overcome Linear Regressions' problem with feature multi-collinearity adding l2 norm regularization of the coefficients to the minimization problem[26].

Random Forest(RF) Non-linear algorithms that rely on ensembles of decision trees with randomness injected into the model in both features and instances selection[10].

We also introduce a meta-algorithm expansion of our own. It is clear that not all influencers should be handled in the same manner and celebrities statistics would show vast differences than those of micro-influencers. We propose a Multiple-Regression model, where data is separated to subsets, in our case, using the K-Means clustering algorithm on the followers statistic [22], and building a regression model for each subset.

Finally, it can be seen in Figures 1b and 1c that the likes and followers' statistics grow in an exponential manner. To handle potential bias towards these features, both in clustering and regression, we transform these statistics using a log scale, i.e., $f(x) = \frac{x}{\ln x}$.

6 EXPERIMENTAL RESULTS

In this section, we present the methodology for evaluating different techniques and introduce two simple yet commonly used baselines. We test our models and present the results of our attempt to measure the influence of Instagram users.

6.1 Methodology

To compare between different models, we employ two commonly used statistics. To test the model's ability to measure influence, we employ the coefficient of determination, denoted R^2. Bound by 1, higher R^2 scores would indicate lower error variance which indicates a tighter model. Comparing the order of the predicted influence with the real influence allows us to rank users. To test the resulting ranking created we use Spearman's rank correlation coefficient, denoted r_s.

To avoid the problems of a model tuned specifically to the test data, we use a five-fold cross validation technique. We randomly split U into five equally sized sets of disjoint Instagrammers and use them as five train-test datasets, each test set contains roughly 20% the size of the original set of users U and the train set is made of the remaining 80%. The results are averaged on the five test cases.

6.2 Baselines

Two natural baselines for measuring influence are to use the user's audience size (followers) or engagement level (number of likes). We use both statistics baselines, utilizing a Linear Regression model.

While outside our scope, for completeness purposes, we used the PageRank extension suggested by Egger [19]. For this, we crawled Instagram, creating a commentators graph around our test users.

6.3 Comparison of Techniques

The results of the R^2 and r_s statistics for the regression models and baselines are provided in Table 1. These results include both clustered and unclustered attempts, as well as, show the result of the feature reduced models.

It is clear that the followers statistic, while intuitive and is often used in real-world scenarios, is the weakest on any given metric. This correlates with previous findings by Cha et al.[15]. The engagement baseline is the best choice for a direct ranking approach as it is almost the best, certainly within error range, and is much simpler to use than the full regression models.

Amongst our suggested models, Multi-Regression was not a useful approach while feature reduction still resulted in strong models with only half the features. When comparing RR and RF, we clearly see that RR is a more accurate model. This is due to a limitation of the RF model - while RR can return any possible value, RF models can return only linear combination of values in the training set and while this result in a better ranker, the predicted value more often overshoots.

Due to resource and time constraints we ran the PageRank algorithm a subset of 10% of the users, resulting in an r_s score of 0.673. These results, only better then the followers baseline, are to be expected given Instagram's flow of information, as discussed in Section 2.

7 CONCLUSIONS AND FUTURE WORK

This work focused on measuring influence and influencer ranking on Instagram, a content sharing OSN. Our definition of influence (Def. 3.1) and the features extracted from public information allowed us to use out-of-the-box regression models to create what is, to our knowledge, the first influence ranking algorithm based on an intuitive score derived from network-oblivious statistics. We have shown general truths regarding Instagram such that the commonly sought out audience size is a poor metric for influence.

Table 1: R^2 and r_s statistics for regression models

	Regression		Multi-Regression	
	R^2	r_s	R^2	r_s
full Ridge Regression	**0.725**	0.848	**0.727**	0.821
full Random Forest	0.626	**0.869**	0.621	**0.861**
minimal Ridge Regression	0.723	0.818	0.727	0.818
minimal Random Forest	0.616	0.864	0.611	0.859
Followers Baseline	0.211	0.757	0.204	0.725
Likes Baseline	0.666	0.859	0.654	0.853

In our work, we did not consider the temporal nature of influence, i.e., the influence of a user is likely to change over time. The rate of change may even depend on the influence itself, as per the rich get richer phenomenon [5].

Lastly, only simple user and posts statistics were used in this work. We believe the use of more complex features would result in stronger models and a better ranking algorithm. These features can be post specific, from the simple "day of the week" to complex "contains faces" [6, 41], user specific, e.g. the user's age or common content type [28, 29], or features relating to a user's audience, such as audience location or age [21, 34].

REFERENCES

[1] Nitin Agarwal, Huan Liu, Lei Tang, and Philip S Yu. 2008. Identifying the influential bloggers in a community. In *Proceedings of the 2008 international conference on web search and data mining*. ACM, 207–218.

[2] Abdullah Al-Bahrani and Darshak Patel. 2015. Incorporating Twitter, Instagram, and Facebook in Economics Classrooms. *The Journal of Economic Education* 46, 1 (2015), 56–67.

[3] Mohammad Shafkat Amin, Baoshi Yan, Sripad Sriram, Anmol Bhasin, and Christian Posse. 2012. Social Referral: Leveraging Network Connections to Deliver Recommendations. In *Proceedings of the Sixth Conference on Recommender Systems (RecSys '12)*. 273–276.

[4] Isabel Anger and Christian Kittl. 2011. Measuring influence on Twitter. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*. ACM, 31.

[5] Camila Souza Araújo, Luiz Paulo Damilton Corrêa, Ana Paula Couto da Silva, Raquel Oliveira Prates, and Wagner Meira. 2014. It is not just a picture: revealing some user practices in instagram. In *Web Congress (LA-WEB), 2014 9th Latin American*. IEEE, 19–23.

[6] Saeideh Bakhshi, David A Shamma, and Eric Gilbert. 2014. Faces engage us: Photos with faces attract more likes and comments on instagram. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 965–974.

[7] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. 2013. Influence-based network-oblivious community detection. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 955–960.

[8] Monica Bianchini, Marco Gori, and Franco Scarselli. 2005. Inside PageRank. *Transactions on Internet Technology (TOIT)* 5, 1 (Feb. 2005), 92–128.

[9] YJ Bijen. 2017. *# AD: The effects of an influencer, comments and product combination on brand image.* Master's thesis. University of Twente.

[10] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[11] Sergey Brin and Lawrence Page. 1998. The Anatomy of a Large-scale Hypertextual Web Search Engine. In *Proceedings of the Seventh International Conference on World Wide Web 7 (WWW7)*. 107–117.

[12] Danny Brown and Sam Fiorella. 2013. *Influence marketing: How to create, manage, and measure brand influencers in social media marketing*. Que Publishing.

[13] Suratna Budalakoti and Ron Bekkerman. 2012. Bimodal Invitation-navigation Fair Bets Model for Authority Identification in a Social Network. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*. 709–718.

[14] Suratna Budalakoti and Ron Bekkerman. 2012. Bimodal invitation-navigation fair bets model for authority identification in a social network. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 709–718.

[15] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. 2010. Measuring user influence in twitter: The million follower fallacy. *Icwsm* 10, 10-17 (2010), 30.

[16] Duanbing Chen, Linyuan Lü, Ming-Sheng Shang, Yi-Cheng Zhang, and Tao Zhou. 2012. Identifying influential nodes in complex networks. *Physica a: Statistical mechanics and its applications* 391, 4 (2012), 1777–1787.

[17] Yuyu Chen. 2016. The rise of 'micro-influencers' on Instagram. (2016).

[18] Victor P Cornet and Natalie K Hall. 2016. Instagram Power Users and their Effect on Social Movements. (2016).

[19] Christopher Egger. 2016. Identifying Key Opinion Leaders in Social Networks-An Approach to use Instagram Data to Rate and Identify Key Opinion Leader for a Specific Business Field. (2016).

[20] Lyon Ethan. 2009. Differing Approaches to Social Influence. (2009). "http://sparxoo.com/2009/10/01/differing-approaches-to-social-influence"

[21] Emilio Ferrara, Roberto Interdonato, and Andrea Tagarelli. 2014. Online popularity and topical interests through the lens of instagram. In *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, 24–34.

[22] Edward W Forgy. 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics* 21 (1965), 768–769.

[23] Thomas Gegenhuber and Leonhard Dobusch. 2017. Making an impression through openness: how open strategy-making practices change in the evolution of new ventures. *Long Range Planning* 50, 3 (2017), 337–354.

[24] Amit Goyal, Francesco Bonchi, and Laks VS Lakshmanan. 2010. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 241–250.

[25] Ashley Ha. 2015. An Experiment: Instagram Marketing Techniques and Their Effectiveness. (2015).

[26] Arthur E Hoerl and Robert W Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12, 1 (1970), 55–67.

[27] Cho-Jui Hsieh, Mitul Tiwari, Deepak Agarwal, Xinyi (Lisa) Huang, and Sam Shah. 2013. Organizational Overlap on Social Networks and Its Applications. In *Proceedings of the 22Nd International Conference on World Wide Web (WWW '13)*. 571–582.

[28] Yuheng Hu, Lydia Manikonda, Subbarao Kambhampati, et al. 2014. What We Instagram: A First Analysis of Instagram Photo Content and User Types.. In *Icwsm*.

[29] Jin Yea Jang, Kyungsik Han, Patrick C Shih, and Dongwon Lee. 2015. Generation like: comparative characteristics in Instagram. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 4039–4042.

[30] Alexy Khrabrov and George Cybenko. 2010. Discovering influence in communication networks using dynamic graph analysis. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*. IEEE, 288–294.

[31] Ayşe Binay KURULTAY. 2012. Dynamics Of Viral Advertising. *The Turkish Online Journal of Design Art and Communication* 2, 2 (2012).

[32] Nadezhda Lisichkova and Zeina Othman. 2017. The Impact of Influencers on Online Purchase Intent. (2017).

[33] Linyuan Lü, Duanbing Chen, Xiao-Long Ren, Qian-Ming Zhang, Yi-Cheng Zhang, and Tao Zhou. 2016. Vital nodes identification in complex networks. *Physics Reports* 650 (2016), 1–63.

[34] Lydia Manikonda, Yuheng Hu, and Subbarao Kambhampati. 2014. Analyzing user activities, demographics, social network structure and user-generated content on Instagram. *arXiv preprint arXiv:1410.8099* (2014).

[35] Robert M Morgan and Shelby D Hunt. 1994. The commitment-trust theory of relationship marketing. *The journal of marketing* (1994), 20–38.

[36] Mira Rakić and Beba Rakić. 2014. VIRAL MARKETING. *Ekonomika* 60, 4 (2014).

[37] Azarias Reda, Yubin Park, Mitul Tiwari, Christian Posse, and Sam Shah. 2012. Metaphor: A System for Related Search Recommendations. In *Proceedings of the 21st International Conference on Information and Knowledge Management (CIKM '12)*. 664–673.

[38] Amanda Richardson, Ollie Ganz, and Donna Vallone. 2013. The cigar ambassador: how Snoop Dogg uses Instagram to promote tobacco use. *Tobacco control* (2013), tobaccocontrol–2013.

[39] Mario Rodriguez, Christian Posse, and Ethan Zhang. 2012. Multiple Objective Optimization in Recommender Systems. In *Proceedings of the Sixth Conference on Recommender Systems (RecSys '12)*. New York, NY, USA, 11–18.

[40] Robert Scoble and Shel Israel. 2005. *Naked conversations: how blogs are changing the way businesses talk with customers*. John Wiley & Sons.

[41] Thiago H Silva, Pedro OS Vaz de Melo, Jussara M Almeida, Juliana Salles, and Antonio AF Loureiro. 2013. A picture of Instagram is worth more than a thousand words: Workload characterization and application. In *Distributed Computing in Sensor Systems (DCOSS), 2013 IEEE International Conference on*. IEEE, 123–132.

[42] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 261–270.

[43] Sven Wilde. 2013. *Viral marketing within social networking sites: the creation of an effective viral marketing campaign*. Diplomica Verlag.

[44] Nathalie Zietek. 2016. Influencer Marketing: the characteristics and components of fashion influencer marketing. (2016).