Network Happiness: How Online Social Interactions Relate to Our Well Being

Johan Bollen and Bruno Gonçalves

1 Introduction

In the normal course of our daily lives we naturally interact with many other individuals: the barista that prepares our daily *venti white chocolate mocha frappuccino*, the bus driver whom we ask for information, the supermarket teller that rings us out, our online acquaintances that we discuss scifi literature with, our coworkers, and our family and loved ones. However, it is clear that not all of these interactions carry the same weight or importance. We may not remember the name of the bus driver or the barista, but we would be remiss if we didn't remember the birthday of our significant other.

In an offline context it is relatively intuitive to observe and distinguish which relationships matter most to us. A small group of people with which we have close personal relations account for most of our social interactions while we dedicate less time or attention to more transactional interactions, such as those with service providers or strangers. Unfortunately for social scientists, it has proven difficult to quantitatively measure the strength and extent of real-world relationships at large scale without intrusive procedures and interventions.

However, our online activities now provide a unique opportunity to conduct such measurements as a by-product of the way in which such systems function. Every *"Like," "retweet,"* or *"mention"* of billions of individuals is recorded and stored in large-scale databases that provide a unique perspective on how individuals interact socially, how they communicate with one another, and which aspects of their

J. Bollen (✉)
School of Informatics and Computing, Indiana University, Bloomington, IN, USA
e-mail: jbollen@indiana.edu

B. Gonçalves (✉)
Center for Data Science, New York University, New York, NY, USA

S. Lehmann, Y.-Y. Ahn (eds.), *Complex Spreading Phenomena in Social Systems*,
Computational Social Sciences, https://doi.org/10.1007/978-3-319-77332-2_14

social lives capture most of their attention [3]. Different systems naturally provide different features, different modes of interaction and, consequently, different views on human social behavior.

In this work we focus on Twitter which is a popular microblogging platform that as of June 30, 2016[1] was used world-wide by over 328 million users. Twitter was designed from the start to allow users to share their content to the world at large in the easiest way possible. As a result, all user-generated Twitter content is public by default and easily accessible through the use of an extensive API,[2] a fact that has long since made Twitter an invaluable resource for academic and industry researchers interested in the study of human behavior, information diffusion, and social network dynamics. Through the Twitter API one is able to easily access both the content user share and their social relations, a feature that makes it particularly suitable for the purposes of studying the relation between individual psychological states and social relationships.

2 Social Interactions

Modern online social network platforms provide a rich set of features and possibilities for users to interact socially. Twitter, in particular, allows users to unilaterally "Follow" another user, "Mention" another user, "Retweet" someone's tweet, or "Like" one another user's tweet. Each of these manners of interaction has a different meaning and, potentially, represents a different type of relationship.

Based on these types of interaction there are different possibilities to decide whether or not two users are in fact "friends."

The simplest method is to define friends as users who regularly engage each other in conversation (via replies). This definition is based on the assumption that the active exchange of information between two parties indicates a social relation. This definition has been used [4] previously by us to empirically verify the well-known Dunbar's number (a cognitive limit on the typical number of active social relationships). While likely corresponding to a "real," offline, relationship, this definition does have the disadvantage of being rather strict; not all friendships involve the active exchange of information through Twitter replies.

Another method which we apply for the rest of this manuscript is to define friendship simply as two users who follow each other, as in Fig. 1. After all, friendship implies a reciprocated, symmetric relation. Celebrities can be followed by thousands and even hundreds of thousands of other users, and might on occasion even reply to messages, but they are not necessarily friends with their Followers, since the relation is not symmetric. While it is rather unlikely that all symmetrical Follow relations correspond to actual friendships, it does provide us with an

[1] https://about.twitter.com/company.

[2] https://dev.twitter.com/rest/public.

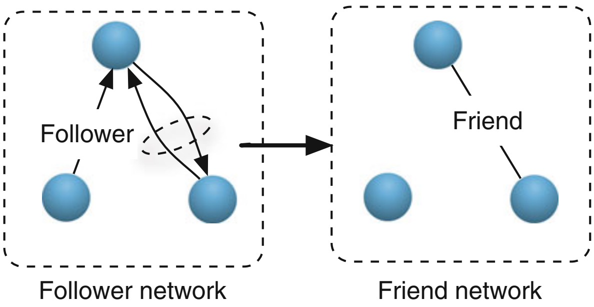

Fig. 1 Friendship ties are by their very nature symmetric, but Twitter connects users by asymmetric Follower relations. This means that Twitter users may Follow other users, but the Follow relationship does not have to be reciprocated. Twitter's Follow relations are thus not sufficient to establish the existence of a Friendship tie between users. In our work, we adopt a minimal definition of a Friendship relation in the Twitter network as two users that share a *reciprocal Follow* relation. This approach does not require additional metadata such as content or frequency of information exchanges and it satisfies the minimum condition that a Friendship tie be symmetric. However, as a result, it does not account for the intensity or degree of the relationship—See [1]

operational definition of online friendship in which information (in the form of tweets) may in principle flow both ways so that each user may potentially influence the other.

The first step of our analysis is to build an empirical friendship network from our Twitter follow data. For this purpose, we collected about 129 million tweets covering the period between November 28, 2008 to May 2009. The API provides us only with a 10% sample of all tweets produced. To avoid issues due to this sampling limitation, we expanded this dataset by retrieving the complete twitter history of all the users in our (sampled) dataset, as well as their follower network. The final mutual Twitter Follower network contains a total of 4,844,430 users (including followers of our users for which we did not collect timeline information).

From this dataset, we eliminate any user that has, on average, less than one tweet per day in the period of our study. In this way, we eliminate spurious users that are unlikely to have a significant impact on their neighbors. Finally, we remove all nonmutual connections to define the friendship network shown in Fig. 2. The giant connected component of our final network has over 102 thousand and a relatively large diameter. Further network statistics can be found in Table 1.

To each edge, we associate a weight, w_{ij}, that measures the social overlap between the two nodes. The overlap is defined as:

$$w_{i,j} = \frac{||C_i \cap C_j||}{||C_i \cup C_j||}, \tag{1}$$

where C_i is the set of friends of node i. Our goal in defining the strength of each connection in this way is twofold: first, this definition of social overlap is purely topological and insensitive to the number of actual interactions between the two

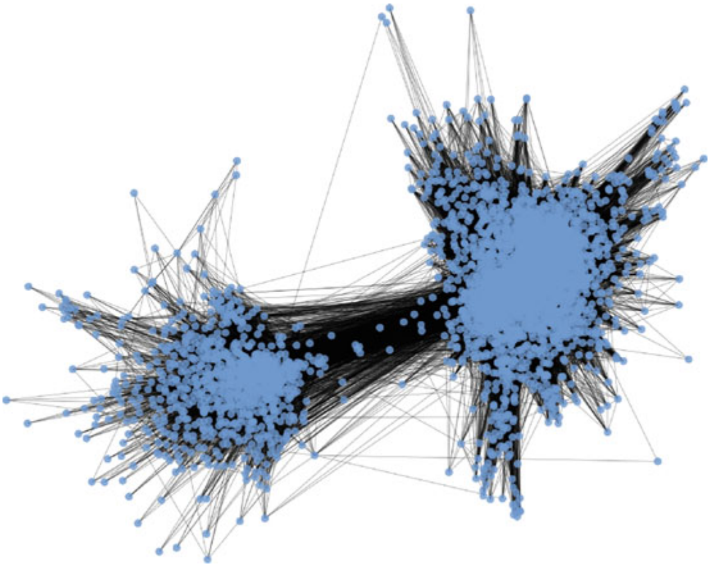

Fig. 2 A force-directed visualization of a sample of the Friendship Network that resulted from our analysis of the Twitter Follow relations between more than 102,000 users. See [1]

users. Second, it gives us a parameter that we may threshold in order to control for shared social context as users with more mutual friends are more likely to be subjected to similar content.

3 Network Structure

We now explore the structure of the friendship network we generated in the previous section. Some fundamental network statistics are listed in Table 1. Particularly significant are the relatively large average degree, $\langle k \rangle = 46.3$ and clustering coefficient, $\langle C \rangle = 0.262$. The high clustering value is typical of real-world social networks [5, 6] with tightly knit groups of friends that are loosely connected through mutual acquaintances.

This type of structure is a result of the definition we used for friendship and helps to explain the relatively large diameter 14 that we observe. Above, we imposed that a link between two individuals is only created if they mutually follow each other. Inside dense friend groups, this happens naturally over the course of repeated interactions and also thanks to the fact that in many cases, these groups are to some degree topical [7]. On the other hand, our strict definition also makes it less likely for us to observe mutual follower relationships between individuals in distant groups, directly increasing the diameter of the network. The full degree distribution can be observed on the right-hand side of Fig. 3. The degree distribution we observe displays a clear broad tailed behavior. This provides further clues to the structure of

Table 1 Network statistics for a Friendship network derived from the Twitter Follow relationships between 102, 009 users

Nodes	102,009
Edges	2,361,547
Density	0.000454
Diameter	14
$\langle k \rangle$	46.300
$\langle C \rangle$	0.262

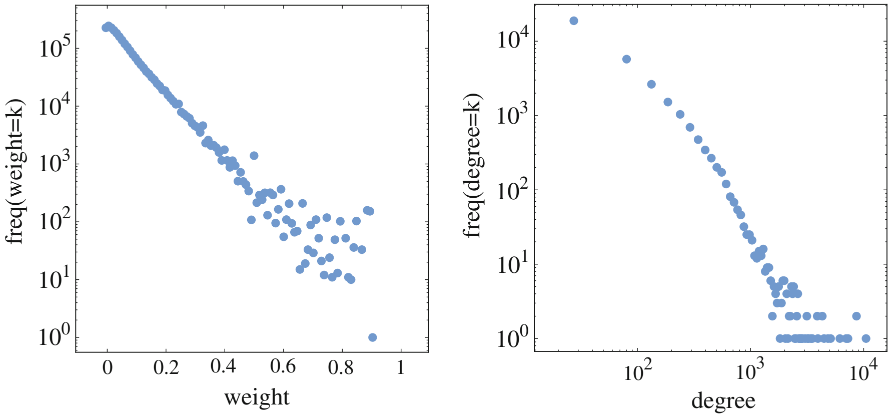

Fig. 3 Distributions of edge weights and node degrees showed considerable skewness indicating that the large majority of connections have low connection weights whereas a few have very high connection weights, and that most users have very few Friendship relations in the network whereas a few individuals having order of magnitude higher number of friends. See [1]

the network as it shows that most nodes have relatively small degrees while a small number of them, the hubs, have collected several thousands of edges and help bind the network together as a whole.

However, not all links are created equal. We assign to each edge a weight corresponding to the number of mutual friends of the two individual at each end of the connection. Naturally, we expect that edges within groups will have higher weights while external edges should correspond to significantly smaller values resulting in a broad tailed weight distribution as shown on the left-hand side of Fig. 3.

4 Friendship Paradox

Hubs, by their very nature, play an important role in maintaining the connectivity of the network. The simple fact that they have such large degree implies that they *must* be connected to nodes in very different locations on the network. However, the picture is even more interesting if we take the opposite perspective, that of the ordinary node that is connected to a hub.

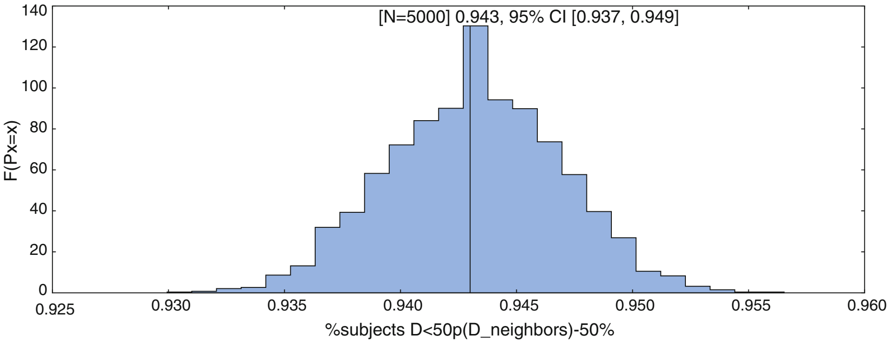

Fig. 4 Histogram of the number of users with lower degree than the median of their friends over 5000 bootstrap realizations. See [2]

As most of us will remember from High School or college, there are advantages to being friends with the most popular kid in school. They know everyone and are better plugged in to the zeitgeist so they can act as brokers of information and introduce us to others we might be interested in meeting. As a result, they have a disproportionately large weight in our lives. This means that we will likely try to connect to them and others like them increasing both their global reach and impact in our lives. If we extend this way of thinking just a couple of steps further we reach a startling conclusion: everyone is trying to connect to these few hubs, resulting in a locally star-like graph. However, we have already observed strong assortativity effects. How can these two phenomena co-exist in the same system?

This observation is indeed paradoxical, but real none the less and is known as the Friendship Paradox: your friends are similar to you, but they also have more friends than you on average [8]. We investigate this paradox by measuring the fraction of users for whom the median degree of their friends is higher than their own. The result of this measurement is a single number of which we have no further information. We study its robustness through a simple bootstrapping procedure. Instead of measuring it over the full network, we evaluate it repeatedly over a small randomly selected fraction of the network. In Fig. 4 we plot the histogram of the fraction of nodes whose degree is smaller than the median of their friends' degrees taken over 5000 bootstrapping procedures. As we can see, our network displays a very strong Friendship Paradox. A large majority of users find themselves less popular than their friends on average.

5 Subjective Well-Being

After the data mining and preparation procedure outlined above, we have the complete Twitter history of all users in our network for a period of 6 months.

We define the Subjective Well-Being of an individual as the average valence (+ or −) of the content produced by him or her. To this end we apply the OpinionFinder (OF)[3] lexicon that assigns a positive (+1) or negative (−1) valence value to a set of 8630 words (2718 positive and 4912 negative words).

The subjective well-being $S(u)$ of user u is then defined as the fractional difference between the number of tweets that contain positive OF terms and those that contain negative terms:

$$S(u) = \frac{N_{+}(u) - N_{-}(u)}{N_{+}(u) + N_{-}(u)}, \tag{2}$$

where $N_{-}(u)$ and $N_{+}(u)$ represent, respectively, the numbers of positive and negative tweets for user u.

After this procedure was applied, each node in our undirected, weighted network, has associated with it the average emotional polarity of the respective user, defined on a scale of [1, +1]. The empirical distribution of SWB is shown in Fig. 5. Despite the fact that the OF Corpus contains almost twice as many negative as positive words, we find a skew in the distribution towards positive SWB values with the positive values displaying an almost symmetrical distribution centered at SWB = 0.2. It is also worth to note that most negative values are close to zero, but how are these nodes connected to one another?

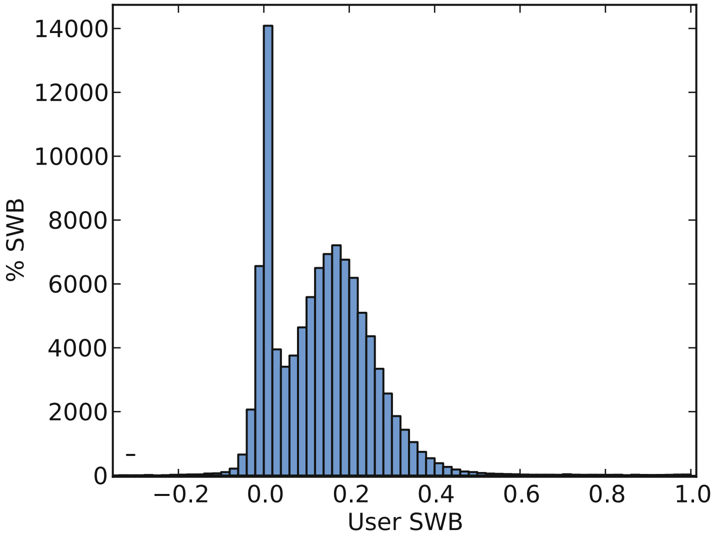

Fig. 5 Distribution of Subjective Well-Being values over all users in our network reveals a strongly bi-modal distribution with two peaks: one slightly below zero and one around SWB = 0.2. See [1]

[3] http://www.cs.pitt.edu/mpqa/opinionfinderrelease/.

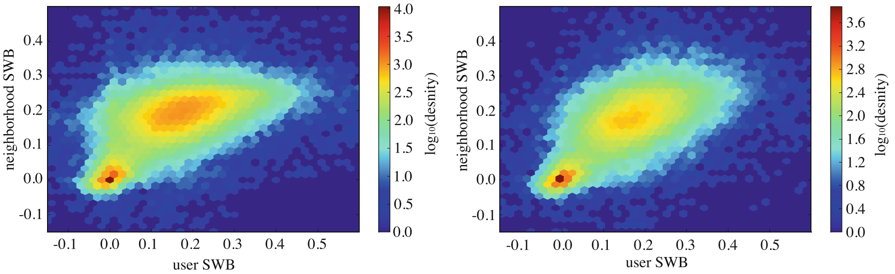

Fig. 6 2D histogram of SWB values for users (x) and their neighborhood (y). Left: all edges included. SWB assortativity = 0.689, N = 102,009 nodes. Right: histogram including only edges with $w_{ij} \geq 0.1$, SBW assortativity = 0.746, N = 59,952. See [1]

We start to answer this question by measuring the correlations between SWB of neighboring users. On the left side of Fig. 6 we plot the $2D$ histogram of the SWB values for users (x) and the average value taken over their neighborhood (y). Surprisingly, we find that this distribution is bi-modal with two clear clusters: one centered around zero and a larger one centered around SWB $\approx$ 0.2. Most points in the figure are located close to the diagonal, indicating a large degree of SWB assortativity. Indeed, we measure the assortativity as the correlation between the two SWB values and find it to be a surprisingly large, namely $R = 0.689$.

However, as we saw in Figs. 3 and 5, a large number of edges have low weight and a large number of nodes have lower values of SWB. Perhaps this is further blurring the results and might explain why we find a SWB cluster near 0? To clarify this possibility we repeat this analysis while keeping only edges with weights $w_{ij} \geq 0.1$ (see [9] for further details). In this way, we are able to keep only the strongest (and thus likely intra-group) edges and help reduce the amount of noise in our results.

The resulting plot is shown on the right-hand side of Fig. 6. The outcome is quite striking. Not only is the second cluster near SWB $=$ 0 still present, but the assortativity has increased significantly to a whopping 0.746 providing strong evidence that these results correspond to significant features of our social system.

6 Happiness Paradox

We finalize our analysis by further considering the correlations between SWB values in the two clusters we found. For this, we divide our users into two groups: a “Happy” group and an "Unhappy" group. The former has high SWB values and is surrounded by friends with equally high SWB value. The latter has low SWB values and so do their friends. This way we compare SWB values only within clusters of comparable individuals.

We use a Gaussian Mixture Model (GMM) to demarcate our Happy and Unhappy groups. The location and distribution of each Gaussian component in the distribution

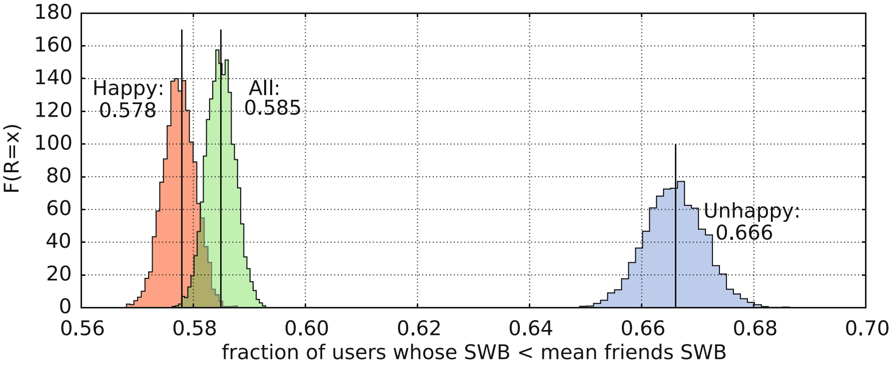

Fig. 7 Distribution of bootstrapped estimates of the magnitude of the Happiness Paradox in our network for the Happy (red) and Unhappy (blue) group, and All (gray). See [2]

of individual vs. mean friend happiness is used to demarcate both groups by simply determining whether the SWB value of a subject and the mean SWB values of their neighbors fall within 2 standard deviations from the center of either one of the components (illustrated by the ellipses in Fig. 7).

Similarly to our Friendship Paradox analysis we also measure how the SWB value of a user is related to that of their friends through a bootstrapping procedure. In Fig. 7 we plot the histograms, taken of 5000 realizations of the bootstrapping procedure of the fraction of users whose SWB is less than the mean of their friends for the full dataset, the “Happy” and the “Unhappy” groups. As we can see, a similar behavior as the one observed for the Friendship Paradox is observed in all three cases: your friends are happier on average than you. We call this the Happiness Paradox.

One particularly interesting feature of these results is the fact that the Unhappy group, despite being the smallest, is the one for which the Happiness Paradox is strongest. This result together with Fig. 6 brings to bear the true strength of this phenomenon. Despite the fact that, in the Unhappy group, you are most likely connected with other Unhappy (SWB < 0) users, they are **still** happier than you. In Fig. 8 we schematically represent the relation between the Friendship and the Happiness Paradoxes.

Finally, we further investigate how the Friendship Paradox manifests itself in the Happy and Unhappy groups as shown in Fig. 9. Subplot A illustrates the boundaries of each group as identified by our GMM approach, while subplots B and C illustrate the friendship paradox for each group. For clarity, we plot the log of the degrees. From this figure, it is clear that both the Friendship and Happiness paradoxes are present and statistically robust in our data set opening up new possibilities of research on the dynamical mechanisms that might help us understand these phenomena.

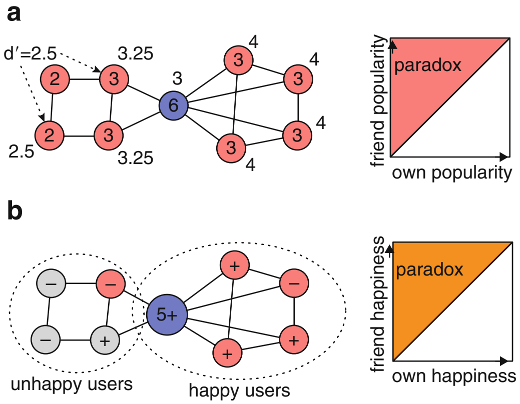

Fig. 8 Diagram that visualizes how in many networks some user are very popular (left) and thereby inflate the average popularity of the Friendship networks they are part of, leading to a Friendship paradox where a majority of users are less popular than their friends on average (red nodes). If Popular individuals are also Happier, their presence in the network will also inflate average Happiness and lead to a Happiness Paradox (right). See [2]. (**a**) Friendship paradox (**b**) Happiness paradox

7 Discussion

The advent of social media has created a unique opportunity to study long-standing questions about how humans form social relations and how these relations affect their well-being, individually as well as collectively. The availability of longitudinal records of what users publish on social media allows us to assess their fluctuating mood state and overall well-being. The records of whom users talk to, about, and whom they follow, provide various perspectives on the multiplex of their social relations. In our work, we adopted an approach that combines a variety of social media data to measure otherwise difficult to quantify social constructs such as "happiness," "well-being," and "friendship" and establish meaningful correlations between how individuals and communities relate to each other and how it may affect their well-being over time. Social media has been in existence for almost a decade establishing records that allow us to study socio-economic phenomena as they emerge and develop over time. This allows the study of longitudinal behavioral and psychological indicators pertaining to individuals and communities over long periods of time.

We caution that many of our results may confound natural phenomena with interface and sample bias. Social media platforms are run by private enterprises that are not bound by requirements to further social science research. Researchers in this field therefore need to carefully consider the potential of self-selection, sample, interface, and social conformity bias in their work. In addition, our work pertains

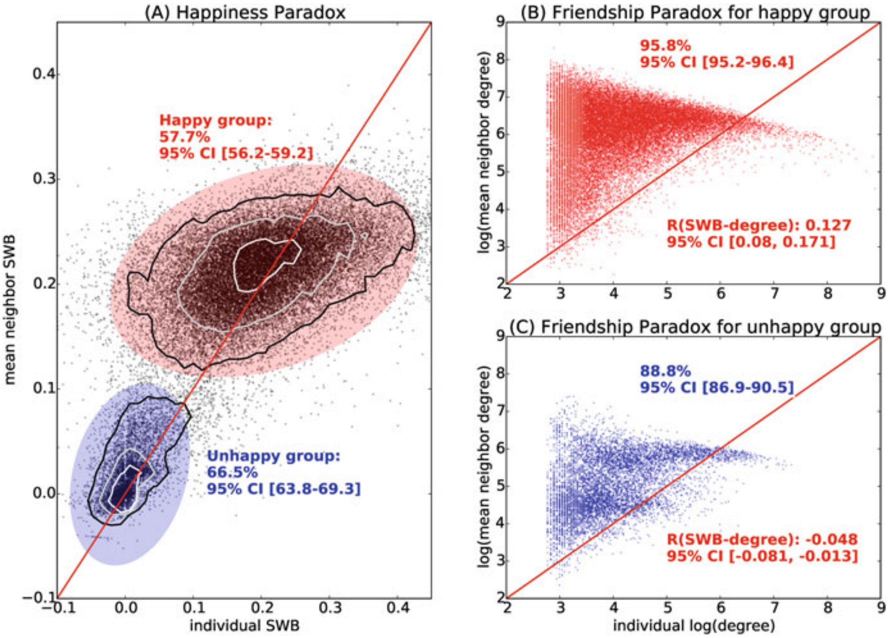

Fig. 9 This graph visualizes the magnitude of the Happiness and Friendship paradox in our sample of Twitter user. (**a**) (left): a majority of users are situated above the diagonal line in which a user's own Subjective Well-Being (SWB) is equal to the mean SWB of the user's friends. In other words, we find a significant Happiness paradox for both Happy and Unhappy groups of users. (**b** and **c**) (right): users in both Happy (red) and Unhappy (blue) groups experience a significant Friendship paradox, i.e. the users find themselves above the diagonal at which their own log(degree) as an indication of Popularity is equal to log(mean degree) of their Friends. See [2]

to snapshots, i.e. data that was harvested in a post-hoc manner and that pertains to specific periods of time in which we were provided access to the data. This situation does not enable controlled experiments and frequently precludes the measurement of "ground truth" with respect to social constructs that are operationalized post-ex-facto in terms of the available data.

In future research, we seek to address these shortcomings. The proliferation of social media platforms may allow the assessment of interface and sample bias as well as the correction of "opportunistic" data harvesting. In addition, we are seeing an increasing trend computational social science of using more traditional social science methods to validate computational indicators derived from social media data to establish "ground truth" and the pre-registration of trials and hypotheses. Our results can be seen as first steps towards an effort to establish a more robust understanding of socio-economic phenomena through the window of large-scale online social networking data.

References

1. Bollen J, Gonçalves B, Ruan G, Mao H (2011) Happiness is assortative in online social networks. Artif Life 17(3):237–251. arxiv:1103.0784, https://doi.org/10.1162/artl_a_00034
2. Bollen J, Gonçalves B, van de Leemput I, Ruan G (2017) The happiness paradox: your friends are happier than you. EPJ Data Sci 6(4). https://doi.org/10.1140/epjds/s13688-017-0100-1
3. Gonçalves B, Perra N (eds) (2015) Social phenomena: from data analysis to models. Springer, Berlin
4. Gonçalves B, Perra N, Vespignani A (2011) Modeling users' activity on twitter: validation of dunbar's number. PLoS One 6:e22656
5. Watts DJ, Strogatz S (1998) Collective dynamics of 'small-world' networks. Nature 393: 440–442
6. Newman MEJ, Park J (2003) Why social networks are different from other types of networks. Phys Rev E 68:036122
7. Aiello LM, Barrat A, Schifanella R, Cattuto C, Markines B, Menczer F (2012) Friendship prediction and homophily in social media. In: ACM transactions on the web (TWEB), vol 6
8. Feld SL (1991) Why your friends have more friends than you do. Am J Sociol 96:1464–1477
9. Bollen J, Gonçalves B, Ruan G, Mao H (2011) Happiness is assortative in online social networks. Artif Life 17:237–251