

# Happiness is assortative in online social networks.

Johan Bollen, Bruno Gonçalves, Guangchen Ruan, & Huina Mao

March 7, 2011

## Abstract

Social networks tend to disproportionately favor connections between individuals with either similar or dissimilar characteristics. This propensity, referred to as assortative mixing or homophily, is expressed as the correlation between attribute values of nearest neighbour vertices in a graph. Recent results indicate that beyond demographic features such as age, sex and race, even psychological states such as “loneliness” can be assortative in a social network. In spite of the increasing societal importance of online social networks it is unknown whether assortative mixing of psychological states takes place in situations where social ties are mediated solely by online networking services in the absence of physical contact. Here, we show that general happiness or Subjective Well-Being (SWB) of Twitter users, as measured from a 6 month record of their individual tweets, is indeed assortative across the Twitter social network. To our knowledge this is the first result that shows assortative mixing in online networks at the level of SWB. Our results imply that online social networks may be equally subject to the social mechanisms that cause assortative mixing in real social networks and that such assortative mixing takes place at the level of SWB. Given the increasing prevalence of online social networks, their propensity to connect users with similar levels of SWB may be an important instrument in better understanding how both positive and negative sentiments spread through online social ties. Future research may focus on how event-specific mood states can propagate and influence user behavior in “real life”.

## 1 Introduction

As the old adage goes, “Birds of a feather flock together”. In network theory, this effect is known as *homophily* [1] or *assortative mixing* [2, 3, 4], occurs in a network when it disproportionately favors connections between vertices with similar characteristics. The opposite trend, that of favouring connections between nodes with different characteristics, is known as disassortative mixing. For example, a friendship network [1] may be highly assortative if it connects individuals who are at similar locations or have similar musical tastes. A heterosexual network [5] on the other hand will be highly disassortative since partners will tend to be of the opposite sex. However, few networks are entirely assortative or disassortative: most will exhibit both properties to some degree depending on the particular characteristic.

Social networks can exhibit significant degrees of assortative mixing with respect to a variety of demographic attributes such as sex, race, age, religion and education, including behavioral and health attributes [6, 7, 8, 9] and even genotypes [10]. Surprisingly, this is also the case for certain psychological states such as loneliness [1]. In the latter case individuals preferentially share relations with individuals who report equally elevated levels of

loneliness and this homophilic tendency increases over time.

Although it is clear that psychological states affect behaviour both online [11] and offline, the mechanisms through which such states exhibit assortativity and contagion across social bonds are not yet fully understood. However, two different processes are conceivable: that individuals seek homophilic social relations to share subjective experiences (homophilic attachment), or that the emotional state of an individual can influence that of the people with which he or she interacts (contagion) [12]. While both possibilities are clearly in play in real-world social interactions, it is not clear whether or not they are present in *online* social systems which do not necessarily emerge from physical contact or in-person communication [13, 14].

The Twitter<sup>1</sup> microblogging service is a case in point. Twitter users can post brief personal updates of less than 140 characters at any time. These updates, known as “tweets”, are distributed to a limited group of “followers”, i.e. other Twitter users who have elected to “follow” the particular user’s tweets [15]. These follower relations are of a fundamentally different nature than their off-line counterparts [16]; they are not necessarily reciprocated, i.e. directed, nor modulated and are mostly focused on the exchange of information. In effect, a Twitter Follower relation simply represents the fact that one individual is interested in the content produced by another, without the requirement that the interest be reciprocated. As a simple example, consider the case of celebrities that attract the attention and interest of a large number of people without reciprocating it. This arrangement results in a social network in the form of a directed, unweighted graph which is quite different from naturally occurring social networks in which friendship ties are generally symmetric and vary in strength. As a consequence, one would expect homophily and assortative mixing of emotional states to be absent or fundamentally altered in online social networking environments, in particular those with asymmetric, unweighted connections such as Twitter.

However, in spite of the expectation that online environments fundamentally alter social interaction, recent results indicate that personal preferences do indeed exhibit homophilic properties in online environments such as BlogCatalog and Last.fm [17]. Tantalizingly this has also been found the case for *sentiment* [18] in LiveJournal<sup>2</sup>. Given the increasing importance of social networking environments in coordinating social unrest [19] and modulating the public’s response to large-scale disasters [20], it has become a matter of tremendous interest whether and how online social networking environments exhibit homophily or even contagion on the level of sentiment and mood and how online tools can be leveraged to gain understanding about social behaviour [21, 22].

Here we investigate whether and to which degree the general happiness or Subjective Well-Being (SWB) [23] of individual Twitter users exhibits assortative mixing. Several previous works have focused on aggregate [24, 25, 26, 27, 28] measurements of mood or emotion in entire communities or systems, but we analyse individual mood state in an online social network. On the basis of a collection of 129 million tweets, we track the SWB levels of 102,009 Twitter users over a 6 months period from the content of their tweets. Each is rated on an emotional scale using a standard sentiment analysis tool. A subsequent assortativity analysis of the Twitter social network then reveals its degree of SWB assortative mixing. Our results indicate that the overall SWB of Twitter users is positive, and highly assortative. In other words, Twitter users are preferentially connected to those with whom they share the same level of general happiness or SWB. Furthermore, tie strength seems to play a significant role in modulating the degree of SWB assortativity.

---

<sup>1</sup>Twitter – <http://www.twitter.com>

<sup>2</sup>LiveJournal: <http://www.livejournal.com/>

## 2 Data and methods

We collected a large set of Tweets submitted to Twitter in the period from November 28, 2008 to May 2009. The data set consisted of 129 million tweets submitted by several million Twitter users. Each Tweet contained a unique identifier, date-time of submission (GMT+0), submission type, and textual content, among other information. Some examples are shown below in Table 1.

ID	date-time	type	text
1	2008-11-28 02:35:48	web	Getting ready for Black Friday. Sleeping out at Circuit City or Walmart not sure which. So cold out.
2	2008-11-28 02:35:48	web	@dane I didn't know I had an uncle named Bob :-P I am going to be checking out the new Flip sometime soon
	...		

Table 1: Examples of Tweet data collected from November 28, 2008 to May 2009 for 4,844,430 users

We complemented this cross-section sample of twitter activity by retrieving the complete history of over 4 million users, as well as the identity of all of their followers. The final Twitter Follower network contained 4,844,430 users (including followers of our users for which we did not collect timeline information). Armed with the social connections and activity of these users we were able to measure the way in which the emotional content of each users varied in time and how it spread across links.

### 2.1 Creating a Twitter “Friend” network

The “Follower” network we collected consists of a directed graph  $G = (V, E)$  in which  $V$  represents the set of all 4,844,430 Twitter users in our collection, and the set of edges  $E \subseteq V^2$  in which each directional edge  $v \in E$  consist of the 2-tuplet  $(v_i, v_j)$  that indicates that user  $v_i$  follows user  $v_j$ . By design the Twitter social network is based on “Follower” relations which are uni-directional and very easy to establish. As such they form a very minimal representation of possible interaction between those who follow and those who are being followed. In fact, it is quite common for a user  $v_i$  to follow a user  $v_j$ , but for  $v_j$  not to follow  $v_i$  back. As such, follower relations are not necessarily indicative of any personal relation which may *de facto* preclude the establishment of assortative mixing and homophily. We therefore distinguish between mere Twitter “Followers” and actual “Friends” [29, 30] by applying the following transformations to the Twitter follower graph  $G$ :

First, we create a network of Twitter Friend relations from the Follower relations in  $G$  by only retaining edges  $(v_i, v_j) \in E$  for which we can find a *reciprocal* relation  $(v_j, v_i)$ , i.e. the set of Friend connections  $E' = \{(v_i, v_j) : \exists(v_j, v_i) \in E\}$ , i.e. two users only share a Friendship tie if they are both following each other.

Second, to exclude occasional users that are not truly involved in the Twitter social network, we only retained those users in our Twitter Friend network that posted more than 1 tweet per day on average over the course of 6 months.

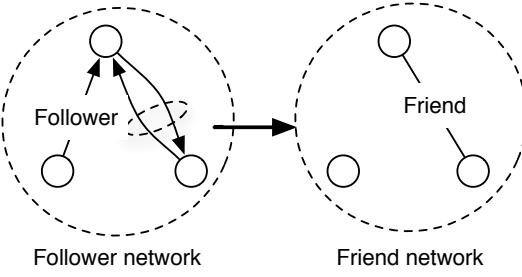


Figure 1: Converting the original Follower network of Twitter into a Friend network by only taking into account reciprocal connections.

Third, we assign a weight  $w_{i,j}$  to each edge  $(v_i, v_j)$  that serves as an indication of the degree to which users  $v_i$  and  $v_j$  have similar sets of friends:

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

where  $C_i$  denotes the neighbourhood of friends surrounding user  $v_i$ . Note that this approach does not take into account the number of tweets exchanged between two users, but simply the degree to which two Twitter users have similar friends.

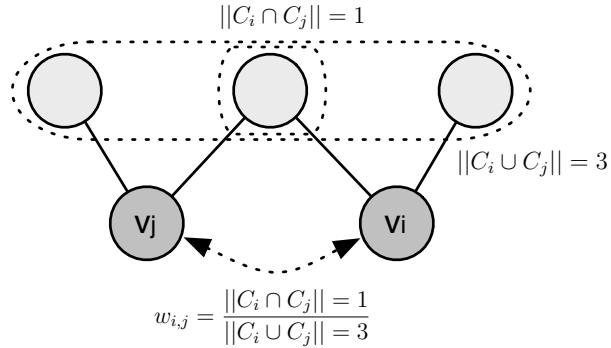


Figure 2: Example of Twitter Friend similarity as calculated according to Eq. 1. Users  $v_i$  and  $v_j$  share 1 friend out of three total. Therefore their connection is assigned a weight  $w_{i,j} = \frac{1}{3}$ .

Finally, we extracted the largest Connected Component ( $G_{CC}$ ) from the resulting network, thereby obtaining a Twitter Friend network of 102,009 users and 2,361,547 edges.

The reduction in nodes from our original Twitter Follower network (4,844,430) to the resulting Friend network (102,009) indicates that in Twitter only a small fraction of users are involved in the type of reciprocated Follower type that we consider indicative of actual social relationships. However, once this reduction has occurred, we

Network parameter	Values
Nodes	102,009 users
Edges	2,361,547 edges
Density	0.000454
Diameter	14
Average Degree	46.300
Average Clustering Coefficient	0.262

Table 2: Network parameters for largest Connected Component of Twitter Friend network.

find that the largest Connected Component of the Friend network,  $G_{CC}$ , retains 97.9% of users in the original Twitter Friend network. This indicates a high degree of connectivity across all users in the final Friend graph. This is further confirmed by the diameter of  $G_{CC}$  which was found to be only 14 in spite of its low density. Other relevant network parameters for  $G_{CC}$  are provided in Table 2.

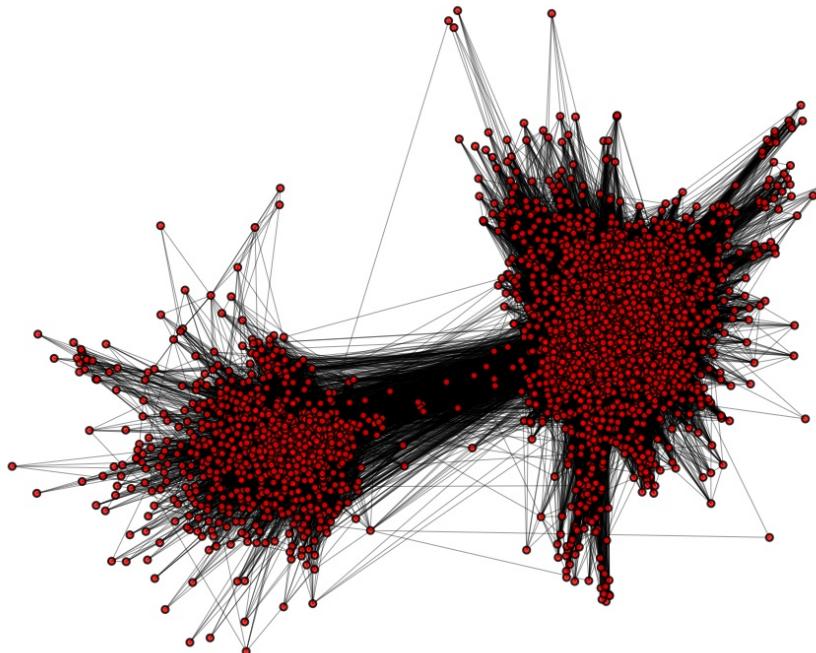


Figure 3: A sub-graph of 3,587 users extracted from the generated Twitter social network (102,009 users and 2,361,547 edges).

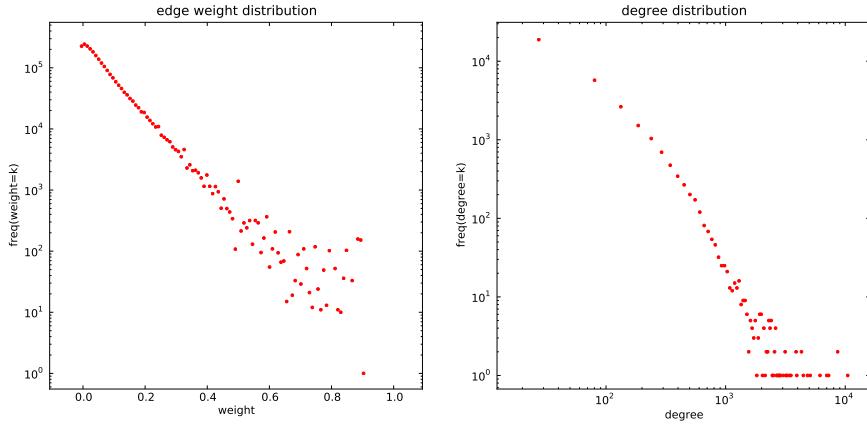


Figure 4: Twitter Friend network edge weight and degree distributions.

Examining the edge weight distribution as shown in Fig. 4 we observe a strongly skewed frequency distribution indicating very many connections in the  $G_{CC}$  with low edge weights ( $w_{i,j} < 0.3$ ) and few connections with very high edge weights ( $w_{i,j} > 0.6$ ). The degree frequency distribution reveals a similar pattern with most users connected to only a few users and a small minority of users connected to thousands of users.

## 2.2 User-level measurements of Subjective Well-Being

We can not directly interrogate Twitter users about their Subjective Well-Being (SWB) [23], but we can infer users' SWB from the aggregate emotional content of their tweets over a period of 6 months. To do so we apply the following procedure.

To reduce noise we only include Twitter users in  $G_{CC}$  that posted at least 1 tweet per day. This guarantees at least 180 tweets for every individual user from which to assess their SWB. We then analyze the emotional content of each user's 6 month record of tweets using OpinionFinder (OF)<sup>3</sup> which is a publicly available software package for sentiment analysis that can be applied to determine sentence-level subjectivity [31]. OF has been successfully used to analyze the emotional content of large collections of tweets [32] by using its lexicon to determine the dominance of positive or negative tweets on a given day. Here we select both positive and negative words that are marked as either "weak" and "strong" from the OF sentiment lexicon resulting in a list of 2718 positive and 4912 negative words. For each tweet in an individual user's 6 month record we count the number of negative and positive terms from the OF lexicon that it contains, and increase the individual user's score of either negative or positive tweets by 1 for each occurrence.

The Subjective Well-Being ( $\mathcal{S}(u)$ ) 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:

$$\mathcal{S}(u) = \frac{N_p(u) - N_n(u)}{N_p(u) + N_n(u)}$$

---

<sup>3</sup><http://www.cs.pitt.edu/mpqa/opinionfinderrelease/>

where  $N_p(u)$  and  $N_n(u)$  represent respectively the number of positive and negative tweets for user  $u$ .

A number of examples is shown in Table 3.

<b>Tweets submitted by high SWB users (<math>&gt; 0.5</math>).</b>
So...nothing quite feels like a good shower, shave and haircut...love it
My beautiful friend. i love you sweet smile and your amazing soul
i am very happy. People in Chicago loved my conference. Love you, my sweet friends
@anonymous thanks for your follow I am following you back, great group amazing people
<b>Tweets submitted by low SWB users (<math>&lt; 0.0</math>).</b>
She doesn't deserve the tears but i cry them anyway
I'm sick and my body decides to attack my face and make me break out!! WTF :(
I think my headphones are electrocuting me.
My mom almost killed me this morning. I don't know how much longer i can be here.

Table 3: Examples of Tweets posted by users with very high and very low SWB values.

## 2.3 Defining SWB assortativity

Having calculated the SWB values of each users, we can now proceed to measure the degree to which the SWB of connected users is correlated. Intuitively, a person can be emotionally influenced by their friends in two, complementary, ways: influence can come from interacting with specific individuals to which one may attribute more importance [33]. We refer to this first type as “pairwise node assortativity” since it assesses the degree to which every two pairwise-connected users have similar SWB values. Another possibility is that each individual is influenced by the overall SWB of all of the people it interacts with. We refer to this second type as “neighborhood assortativity”.

Fig. 5 illustrates this distinction; it shows the actual neighborhood Friend network of an individual in  $G_{CC}$  who has very high SWB values. Nodes are colored according to their SWB values with red indicating high SWB values, blue indicating low SWB values and white indicating neutral or zero SWB values. The particular individual with high SWB values is connected to a local network of equally high SWB individuals (red). The individual could thus be said to be *neighborhood assortative* within this cluster. However, the individual is also connected to several individuals with low SWB values (blue). For each individual connection this is a case of *pairwise disassortativity*. The cluster of low SWB individuals on the other hand exhibits neighborhood assortativity for low SWB values, and the network in its entirety, including both low and high SWB clusters, exhibits strong SWB assortativity; nodes with similar low or high SWB values tend to be connected (blue and red clusters).

We formally define *Pairwise* SWB assortativity as follows: For each edge  $(v_i, v_j)$  in the  $G_{CC}$  of our social network, we extract the corresponding two SWB values, one for the source node and one for the target node. These values are then aggregated into two vectors,  $\mathcal{S}(S)$  and  $\mathcal{S}(T)$  for sources and targets respectively. The value

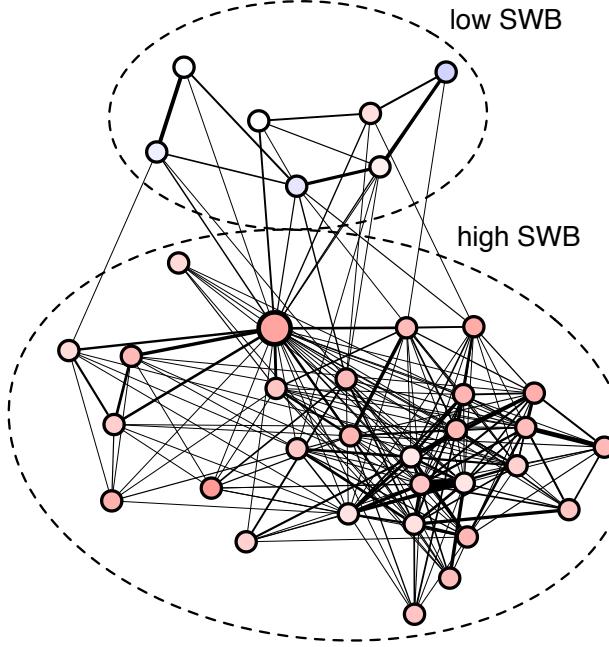


Figure 5: Neighborhood network of a very high SWB individual (center). Blue, white, and red node colors correspond respectively to low, neutral and high SWB values.

of the pairwise assortativity, denoted  $A_P(G_{CC})$ , is then given by the Pearson correlation coefficient  $\rho$  of these two vectors.

$$A_P(G_{CC}) \equiv \rho(\mathcal{S}(S), \mathcal{S}(T)) = \frac{1}{n-1} \sum_i \left[ \left( \frac{\mathcal{S}(S_i) - \langle \mathcal{S}(S) \rangle}{\sigma(\mathcal{S}(S))} \right) \left( \frac{\mathcal{S}(T_i) - \langle \mathcal{S}(T) \rangle}{\sigma(\mathcal{S}(T))} \right) \right] \quad (2)$$

The pairwise assortativity is then defined in the  $[-1, +1]$  interval, with  $-1$  indicating perfect disassortativity,  $0$  indicates a lack of any assortativity, and  $+1$  meaning perfect assortativity.

The *neighborhood* assortativity of  $G_{CC}$  with regards to SWB, denoted  $A_N(G_{CC})$  can be calculated as follows.

For each user  $u \in V$ , we define its neighborhood:

$$\kappa(u) = \{\forall v : \exists(u, v) \in E\} \quad (3)$$

so that  $\kappa(u)$  or  $\kappa_u$  represents the set of users that user  $u$  is connected to. We then calculate an average SWB value for  $\kappa(v)$  which we denote

$$\overline{\mathcal{S}(\kappa_u)} = \frac{1}{\|\kappa(u)\|} \sum_{v \in \kappa(u)} \mathcal{S}(v) \quad (4)$$

We can now define two vectors, one for the SWB values of every unique user and one for the average SWB values of their neighborhoods, denoted by  $\mathcal{S}(U)$  and  $\overline{\mathcal{S}(\kappa)}$ . The neighborhood assortativity of the network  $G_{CC}$  with regards to SWB, denoted  $A_\kappa$ , is then given by the correlation function  $\rho$  computed over these two vectors as follows:

$$A_K(G_{CC}) \equiv \rho(\mathcal{S}(U), \overline{\mathcal{S}(\kappa)}) = \frac{1}{n-1} \sum_u \left[ \left( \frac{\mathcal{S}(u) - \langle \mathcal{S}(U) \rangle}{\sigma(\mathcal{S}(U))} \right) \left( \frac{\overline{\mathcal{S}(\kappa_u)} - \langle \overline{\mathcal{S}(\kappa)} \rangle}{\sigma(\overline{\mathcal{S}(\kappa)})} \right) \right] \quad (5)$$

with the sum to be taken over every user,  $u$ .  $A_K(G_{CC})$  then represents the correlation between the SWB values of user  $v_i$  and the mean SWB values of its Friends. Similarly to the pairwise version,  $A_{\kappa_i}$  is expressed in the range  $[-1, +1]$  where  $-1$  indicates perfect neighborhood disassortativity and where  $+1$  indicates perfect neighborhood assortativity.

## 3 Results and discussion

### 3.1 SWB distribution

In Figure 6 we plot the probability distribution of Subjective Well-Being values across all Twitter users in our sample. The distribution seems bimodal with two peaks: one in the range  $[-0.1, 0.1]$  and another in the range  $[0.2, 0.4]$ . Excluding users whose SWB=0 (due to a lack of emotional content in their Tweets) we find that a majority of Twitter users in our sample have positive SWB values in a rather narrow range  $[0.1, 0.4]$  with a peak at SWB=0.16. This is confirmed by the cumulative distribution shown on the left-bottom of Fig. 6; 50% of users have SWB values  $\leq 0.1$ , and 95% of users have SWB values  $\leq 0.285$ .

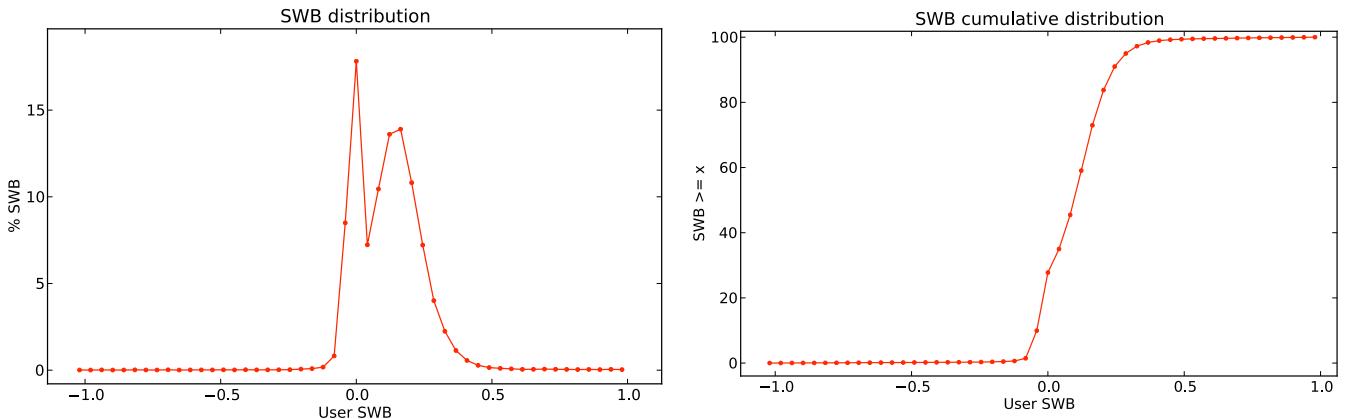


Figure 6: Probability distribution (%) and cumulative distribution (%) of Subjective Well-Being (SWB) and Emotionality values for our sample of Twitter users.

### 3.2 Pairwise and neighborhood SWB assortativity

In Eq. 2 we defined pairwise SWB as the correlation between the SBW values of connected users in our Twitter Friends networks, whereas Eq. 5 defined neighborhood assortativity was defined as the correlation between the SWB values of individual users and the mean SWB values of their neighbors in the graph  $G_{CC}$ .

The assortativity values were found to be  $0.443^{***}$  ( $N=2,062,714$  edges) for the pairwise SBW assortativity and  $0.689^{***}$  ( $N=102,009$ )<sup>4</sup> for the neighborhood assortativity. Both correlations are highly statistically significant ( $p$ -values  $< 0.001$ ) for the sample sizes.

Regarding pairwise SWB assortativity, the scatterplots on the left of Fig. 7 and Fig. 8 show the distribution of SWB values across the sample of all edges and nodes in  $G_{CC}$  and confirm the observed correlation between the SWB values of connected or neighboring users in  $G_{CC}$ .

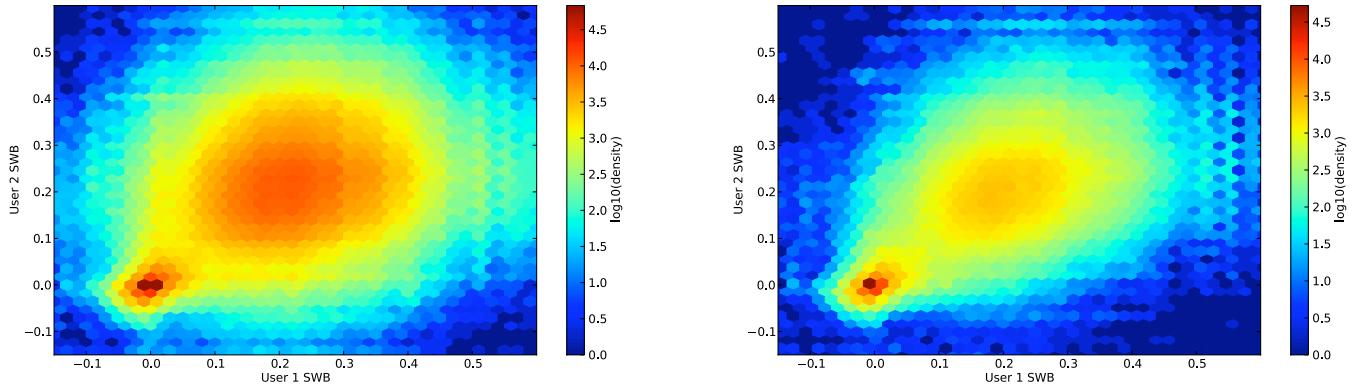


Figure 7: Scatterplot of SWB values for user connected in Twitter Friends network. Left: all edges included. SWB assortativity= $0.443^{***}$ ,  $N=2,062,714$  edges. Right: scatterplot includes edges  $w_{i,j} \geq 0.1$ , SWB assortativity= $0.712^{***}$ ,  $N=479,401$

The pairwise assortativity scatterplot (Fig. 7-left) indicates a significant amount of scatter, commensurate to the lower correlation value of  $0.443^{***}$  which is nevertheless statistically highly significant. The observed relation is not obviously linear. The distribution of values is affected by the bimodal distribution of SWB values as shown in Fig. 6; large numbers of observations cluster at SWB values within the ranges of either  $[-0.05, 0.05]$  and  $[0.1, 0.3]$ . The clustering pattern of Fig. 7 however indicates that users with SWB values in a particular range are preferentially connected to users within that same range, thereby confirming the observed positive pairwise SWB assortativity.

The neighborhood assortativity scatterplot (Fig. 8-left) indicates a similar effect but here the clustering of users is less pronounced and the amount of scatter is lower than that observed for the pairwise assortativity scatterplot,

<sup>4</sup>The sample sizes for pairwise assortativity and neighborhood assortativity) are expressed in edges and nodes respectively, since the former correlation is calculated on the basis of a sample of edges that connect pairs of nodes whereas the other is calculated on the basis of a sample of nodes and their neighborhood

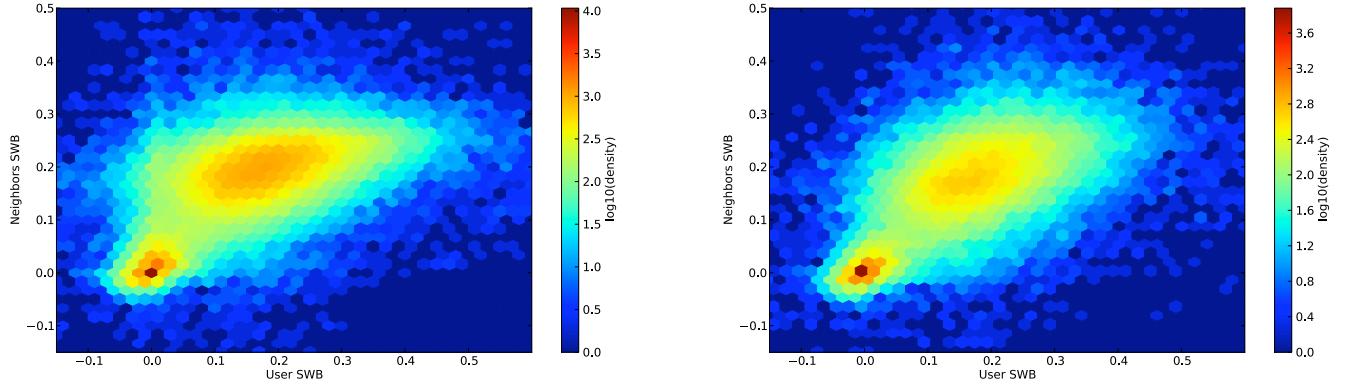


Figure 8: Scatterplot of SWB values for users (x) and their neighborhood (y) connected in Twitter Friends network. Left: all edges included. SWB assortativity=0.689\*\*\*, N=102,009 nodes. Right: scatterplot includes edges  $w_{i,j} \geq 0.1$ , SBW assortativity=0.746\*\*\*, N=59,952

commensurate to the higher neighborhood assortativity value of 0.689\*\*\*. Although less pronounced the bimodal distribution of SWB values is apparent and leads to a clustering of user and neighborhood SWB values in the ranges of  $[-0.05, 0.05]$  and  $[0.1, 0.3]$ . Nevertheless it is again the case that users with SWB values in either range are most likely to be connected to users or neighborhoods with SWB values in the same range. The distribution of user and neighborhood SWB value is furthermore in line with a positive, linear relationship.

### 3.3 Edge weight and SWB assortativity

Pairwise SWB assortativity and neighborhood SWB assortativity diverge significantly ( $0.443^{**} < 0.689^{***}$ ). The former is based on the pairwise comparison of SWB values across all connections in  $G_{CC}$  many of which may be weak or irrelevant connections from the perspective of indicating actual Friend ties. To measure the effect of edge weights, we calculate pairwise and neighborhood assortativity values under different edge thresholds, i.e. we only take into account edges in  $G_{CC}$  whose weight as defined in Eq. 1 is  $w_{i,j} \geq \epsilon$  where  $\epsilon \in [0, 1]$  represents a given edge threshold. The consequent assortativity calculations will therefore more strongly reflect only those connections between users that are indicative of stronger Friend relations (higher  $w_{i,j}$ ). In other words we are verifying whether stronger user relations lead to higher or lower assortativity.

The results of the calculation of pairwise and neighborhood SWB assortativity under various edge thresholds are shown in Table 4 and visualized in Fig. 9. Values for  $\epsilon > 0.8$  are excluded since the correlation coefficients were not statistically significant ( $p\text{-value} < 0.1$ ). The graph in Fig. 9 overlays the different pairwise and neighborhood SWB assortativity values along with the number of remaining edges and nodes under the given edge threshold, i.e. the sample size for the given assortativity calculation.

Both pairwise and neighborhood assortativity values increase as the edge threshold  $\epsilon$  increases, but not in a linear manner. Pairwise SWB assortativity values increase sharply as  $\epsilon$  increases from 0 to 0.10 and afterwards stabilizes at a value of approximately 0.750 which is maintained in the interval  $\epsilon \in [0.15, 0.85]$ . In other words, removing

Edge threshold ( $\epsilon$ )	Pairwise		Neighborhood	
	A(SWB)	N edges	A(SWB)	N nodes
0.0	0.443***	2,062,714	0.689***	102,009
0.10	0.712***	479,401	0.746***	59,952
0.20	0.754***	128,261	0.769***	33,693
0.30	0.755***	36,255	0.780***	16,334
0.40	0.743***	10,355	0.779***	7,699
0.50	0.757***	3,255	0.781***	3,793
0.60	0.798***	1,375	0.805***	1,439
0.70	0.755***	689	0.816***	502
0.80	0.434***	301	0.768***	149
0.90	-	-	-	-

Table 4: Pairwise and Neighborhood Subjective Well-Being assortativity values  $A(SWB)$  vs. edge threshold  $\epsilon$ .  
\*\*\*: p-value < 0.001.

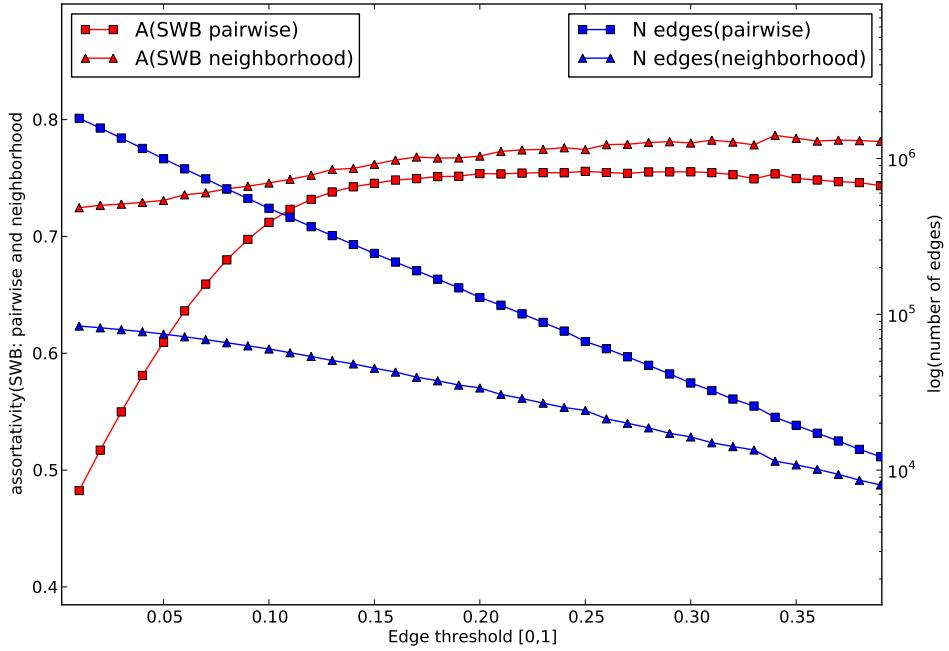


Figure 9: Pairwise Subjective Well-Being assortativity and  $\log(\text{number of edges})$  values vs. edge weight thresholds.

edges with  $w_{i,j} < 0.1$  increases pairwise SWB assortativity considerably, but the removal of edges with higher  $w_{i,j}$  values has little to no additional effect. We observe that  $\epsilon = 0.1$  reduces the number of edges by a fifth,

namely from 2,062,714 to 479,401 indicating that a large number of edges are characterized by low similarity values which when included lead to lower pairwise SWB assortativity.

The neighborhood SWB assortativity increases with higher  $\epsilon$  values but in a less pronounced manner. At  $\epsilon = 0$  we find a neighborhood SWB assortativity value of 0.689 which increases to approximately 0.760 for  $\epsilon \in [0.10, 0.90]$ . We observe significant declines in the number of nodes that remain under increasing  $\epsilon$  values as was the case for pairwise assortativity.

At a value of  $\epsilon = 0.1$  we find the highest pairwise and neighborhood SWB assortativity values combined with the largest sample sizes, excluding no threshold i.e.  $\epsilon = 0$ . We therefore generate new scatterplots of SWB values for pairwise and neighborhood SWB assortativity at  $\epsilon = 0.1$  as shown in Fig. 7 (right) and Fig. 8 (right). The respective scatterplots reflect higher assortativity values; we find less scatter, a stronger positive and linear relation between SWB value of connected users, and a less pronounced clustering caused by the bimodal distribution of SWB values.

### 3.4 Discussion

The above outlined results indicate the following.

First, Twitter users in general exhibit a low to moderate SWB, with very few users being characterized by low SWB values. The Twitter population in our sample can therefore in the mean be considered moderately happy. This observation is most likely an underestimation given the relative preponderance of negative terms in the OF lexicon. However, the SWB distribution is bi-modal showing a group of Twitter users with zero to very low SWB values, i.e. those that are on the average somewhat happy, and another group with more pronounced, higher SWB values. This may result from socio-cultural differences in how emotions and mood are expressed on Twitter. Some users may infrequently express their emotional states whereas other are more prone to do so.

Second, we find statistically significant levels of pairwise and neighborhood SWB assortativity indicating that Twitter users either prefer the company of users with similar SWB values (homophilic attachment) or converge on their Friends' SWB values (contagion). The relation between user SWB values is not linear and biased by the bi-modal distributions of SBW values causing users to be clustered in two groups with equally low or high SWB values. In other words, low SWB users are connected to low SWB and high SWB users are connected to high SWB users. Again, this may confirm the notion that distinct socio-cultural factors affect the expression of emotion and mood on Twitter, and cause users to cluster according to their degree of expressiveness as well as SWB.

The results of measuring pairwise and neighborhood assortativity under different edge weights indicate that we find a stronger and more significant relation between the SWB values of connected users when we only take into account connections with higher  $w_{i,j}$  weights, i.e. those that are deemed more reliable indicators of actual Friend ties. A possible mechanism to explain the difference in magnitude between pairwise and neighborhood assortativity might be that users' neighborhoods contain individuals that they are indeed strongly assortative with, and whose SWB values affect mean neighborhood SWB values and thus neighborhood assortativity overall, whereas

they are “drowned out” in the process of making pairwise comparisons between all individuals that a user is connected to in the process of calculating pairwise SWB assortativity.

For example, users might generally have 10 neighbors, but are generally highly SWB assortative with only 1 Friend. In the calculation of pairwise assortativity, this leads to 10 pairwise comparisons between SBW values only one of which contributes to the overall observed pairwise SWB assortativity in the graph. However, the neighborhood assortativity relies on a mean SWB value calculated for the entire neighborhood, including the 1 highly assortative individual. The latter thus influences the average SBW value for the entire neighborhood causing an increased neighborhood assortativity value.

The greatest improvement in assortativity values indeed occurs for pairwise SWB assortativity which is most affected by the preponderance of weakly weighted connections since it is defined at the level of all individual user to user connections. Both pairwise and neighborhood SWB assortativity converge on a value of approximately 0.750 which indicates a significant degree of SWB assortativity in our Twitter Friend graph  $G_{CC}$ .

## 4 Conclusion

Recent findings show that assortative mixing can occur in a variety of social contexts and personal attributes. Here we show that Subjective Well-Being is equally assortative in the Twitter social network, i.e. the SWB of individuals that have reciprocal Twitter follower links are strongly related. Happy users tend to connect to happy users whereas unhappy users tend to be predominantly connected to unhappy users. The convergence of pairwise and neighborhood assortativity under increasing edge weight thresholds indicates that users tend to be most assortative with a limited number of individuals that they have strong social ties to and that weaker ties fulfill a different social role possibly as outlined by Centola et al (2007)[34].

We do not address the social or cognitive mechanisms that cause the observed SWB assortativity. Two different mechanisms may be at work[35]. The first is based on the notion of “homophily”, i.e. users and connections tend to preferentially connect to users with similar SWB values. As an online social network grows, new connections are thus biased towards connecting individuals with similar SWB values. This process may be modeled in terms of “preferential attachment” theory. The second mechanism that may cause SWB assortativity is that of “mood contagion”, namely that connected users converge to similar SWB values over time. In other words, being connected to unhappy users can make one unhappier and vice versa. The latter suggests that users may control their own level of SWB by choosing the right set of online friends and influence their Friends’ SWB by creating strong social ties and hoping for some form of SWB contagion to take place. A third possibility is that users assess or express their SWB relative to that of their friends. As a user’s neighborhood becomes happier, this may affect their own expression of SWB-related sentiment. This phenomenon may occur at the level of entire cultures which may be comparatively more or less prone to open expressions of individual sentiment.

At this point our research does not offer any information on which of these mechanisms cause the observed SWB assortativity or in fact whether both may be occurring. Future research will therefore focus on analyzing user connections and SWB values over time, and relating these changes in the framework of homophily and preferential attachment[36]. Twitter has now become a major international phenomenon, and this investigation must therefore include linguistic, cultural and geographic factors.

## Acknowledgements

We are grateful to Eliot Smith, Peter Todd and Ishani Banerji for their very useful feedback and input throughout the research that led to this paper. This research was supported by NSF Grant BCS #1032101.

## References

- [1] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1):415–444, August 2001.
- [2] Mark E J Newman. Assortative mixing in networks. *Phys. Rev. Lett.*, 89:208701/1–4, 2002.
- [3] Mark E J Newman. The structure and function of complex networks. *SIAM Review*, 45:167, 2003.
- [4] Mark E J Newman. Mixing patterns in networks. *Physical Review E*, 67(2):26126, 2003.
- [5] Luis E. C. Rocha, Fredrik Liljeros, and Petter Holme. Information dynamics shape the sexual networks of internet-mediated prostitution. *PNAS*, 107:5706, 2010.
- [6] Herminia Ibarra. Homophily and Differential Returns: Sex Differences in Network Structure and Access in an Advertising Firm. *Administrative Science Quarterly*, 37(3):422, September 1992.
- [7] Kelly a. Mollica, Barbara Gray, and Linda K. Trevino. Racial Homophily and Its Persistence in Newcomers' Social Networks. *Organization Science*, 14(2):123–136, March 2003.
- [8] Nicholas a Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *The New England journal of medicine*, 357(4):370–9, July 2007.
- [9] John T Cacioppo, James H Fowler, and Nicholas A Christakis. Alone in the Crowd: The Structure and Spread of Loneliness in a Large Social Network. *Journal of personality and social psychology*, 97(6):977–991, 2010.
- [10] James H. Fowler, Jaime E. Settle, and Nicholas A. Christakis. Correlated genotypes in friendship networks. *PNAS*, 108:1993, 2011.
- [11] Anna Chmiel, Julian Sienkiewicz, Georgios Paltoglou, Kevan Buckley, Mike Thelwall, and Janusz A. Holyst. Negative emotions boost users activity at bbc forum. Technical Report 1011.5459, arXiv, 2010.
- [12] Brian Parkinson and Gwenda Simons. Affecting others: social appraisal and emotion contagion in everyday decision making. *Personality and social psychology bulletin*, 35(8):1071–84, August 2009.
- [13] Winter a Mason, Frederica R Conrey, and Eliot R Smith. Situating social influence processes: dynamic, multidirectional flows of influence within social networks. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc*, 11(3):279–300, August 2007.

- [14] Jeffrey R Huntsinger, Janetta Lun, Stacey Sinclair, and Gerald L Clore. Contagion without contact: anticipatory mood matching in response to affiliative motivation. *Personality and social psychology bulletin*, 35(7):909–22, July 2009.
- [15] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web mining and Social Network Analysis*, pages 56–65, New York, NY, USA, 2007. ACM.
- [16] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 591—600, Raleigh, North Carolina, USA, 2010. ACM.
- [17] Halil Bisgin, Nitin Agarwal, and Xiaowei Xu. Investigating Homophily in Online Social Networks. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 533–536, Toronto, Canada, August 2010. IEEE.
- [18] Reza Zafarani, William D Cole, and Huan Liu. Sentiment Propagation in Social Networks : A Case Study in LiveJournal. In Patricia Chai, Sun-Ki And Salerno, John And Mabry, editor, *SBP 2010 - Advances in Social Computing, Lecture Notes in Computer Science (6007)*, pages 413–420. Springer Berlin, 2010.
- [19] Alina Mungiu-Pippidi and Igor Muntean. Moldova’s “Twitter” Revolution. *Journal of Democracy*, 20(3), 2009.
- [20] Onook Oh, Kyounghee Hazel Kwon, and H. Raghav Rao. An exploration of social media in extreme events: rumor theory and twitter during the Haiti earthquake. In *International Conference on Information Systems (ICIS)*, page Paper 231, St. Louis, Missouri, USA, 2010.
- [21] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. *Science*, 323:5915, 2009.
- [22] Sang Hoon Lee, Pan-Jun Kim, Yong-Yeol Ahn, and Hawoong Jeong. Googling social interactions: Web search engine based social network construction. *PLoS One*, 5:e11233, 2010.
- [23] Ed Diener. *Subjective Well-Being*, pages 11–58. Springer Netherlands, 2009.
- [24] K. Balog, G. Mishne, and M. de Rijke. Why are they excited? identifying and explaining spikes in blog mood levels. In *11th Conference of the European Chapter of the Association for Computational Linguistics. Trento, Italy*, 2006.
- [25] Peter Sheridan Dodds and Christopher M Danforth. Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents. *Journal of Happiness*, 11(4), July 2009.
- [26] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist. Pulse of the nation: visualizing the mood of twitter. <http://www.ccs.neu.edu/home/amislove/twittermood/>, 2010.
- [27] Johan Bollen, Huina Mao, and Alberto Pepe. Determining the public mood state by analysis of microblogging posts. In *Proceedings of the Proc. of the Alife XII Conference*, Odense, Denmark, 2010. MIT Press.

- [28] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, In press, 2011.
- [29] M. Szell, R. Lambiotte, and S. Thurner. Multirelational organization of large-scale social networks in an online world. *Proceedings of the National Academy of Sciences*, 107(31), July 2010.
- [30] B A Huberman, D M Romero, and F Wu. Social networks that matter: Twitter under the microscope. *First Monday*, 14:1, 2008.
- [31] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT05*, (October):347–354, 2005.
- [32] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, Washington DC, USA, 2010. AAAI Press.
- [33] Jukka-Pekka Onnela and Felix Reed-Tsochas. Spontaneous emergence of social influence in online systems. *PNAS*, 107:18375, 2010.
- [34] Damon Centola and Michael Macy. Complex Contagion and the weakness of long ties. *American Journal of Sociology*, 113(3):702–734, 2007.
- [35] Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21544–9, December 2009.
- [36] Cosma Rohilla Shalizi and Andrew C. Thomas. Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. *Sociological Methods and Research (arXiv:1004.4704v3)*, 2011.