Assortativity of suicide-related posting on Twitter

by

Ian Cero

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 4, 2018

Committee

Tracy K. Witte, Chair, Associate Professor of Psychology
Barry Burkhart, Professor of Psychology
Chris Correia, Professor of Psychology
John Rapp, Associate Professor of Psychology

Abstract

Networks in which similar individuals are more likely to associate with one another than their dissimilar counterparts are called *assortative*. Such patterns are a hallmark of human social networks, in which numerous phenomena (e.g., mood, weight) cluster among like-type members. The clustering of suicides in time and space implies such fatalities likely also have socially assortative features, and suggests other forms of suicide-related behavior may as well. This investigation evaluated the assortativity of suicide-related verbalizations (SRV) by machine coding 64 million posts from 17 million unique users of the Twitter social media platform, collected over two distinct 28-day periods. These data were used to assemble a network, in which users were defined as socially linked if they mutually replied to each other at least once. Bootstrapping revealed that SRV was significantly more assortative than chance, up to at least six degrees of separation (i.e., people six links apart were still more similar on SRV than chance). When user mood was controlled, SRV assortativity still remained significantly higher than chance through two degrees of separation, indicating this effect was not just an artifact of mood. Discussion demonstrates how exploiting assortative patterns can improve the efficiency of suicide risk detection.

Table of Contents

Introduction

Suicide is a leading cause of death in the United States and around the world, claiming roughly 800,000 global lives each year (Drapeau & McIntosh, 2015; World Health Organization [WHO], 2012). In addition to the direct loss of life, it is accompanied by substantial emotional and economic burden to millions of surviving loved ones and to broader society (Goldsmith, Pellmar, Kleiman, & Bunney, 2002). However, unlike many other leading causes of death, suicide is not a disease state with well-defined pathological mechanisms (Knox, Conwell, & Caine, 2004). It is instead the fatal outcome of a complex system of interacting individual, social, and environmental factors (Yip et al., 2012), for which the utility of traditional data sources and analytic frameworks are increasingly hamstrung (Christensen, Cuijpers, & Reynolds, 2016; Franklin, Ribeiro, et al., 2016; Knox et al., 2004).

Indeed, recent meta-analytic evidence indicates that predictive estimates from longitudinal studies of suicidal behavior are ultimately too weak for use in clinical settings, even when based on the strongest predictors available (i.e., a known history of self-injurious thoughts and behavior; Ribeiro et al., 2016). Equally concerning, the accuracy of such predictive estimates has not improved to a statistically significant degree in 50 years, suggesting this limitation is unlikely to resolve without an active change in research strategy (Franklin, Ribeiro, et al., 2016). Alongside these research deficits, the rate of death by suicide has not meaningfully decreased in the United States for several decades (Centers for Disease Control and Prevention [CDC], 2016), underscoring the need to explore both novel data and models to inform prevention (Christensen et al., 2016). In response to that call, this investigation departs from previous work focused exclusively on either individual-level risk or population-level risk. We instead investigate features of suicidal behavior at the level of the social network, thus moving toward a quantitative reconciliation of group-level risk patterns and the individual-level factors that produce them.

Assortativity and suicidal behavior

Social clustering is an important feature of suicidal behavior in humans (Gould, Wallenstein, & Davidson, 1989), causing significant concern in the communities where it is perceived (Haw, Hawton, Niedzwiedz, & Platt, 2013). Research has identified both different types of clusters and different mechanisms by which it might occur (Joiner, 1999). For example, atypical elevations in the rate of suicide have been described in narrow time-frames (e.g., after mass media exposure; Gould, 1990), narrow geographic spaces (e.g., underground trains; Farmer, O'Donnell, & Tranah, 1991), as well as narrow temporal-geographic spaces (e.g., several suicides during a short period in a single school; Brent et al., 1989). Non-fatal suicide attempts have also been shown to cluster (Gould, Petrie, Kleinman, & Wallenstein, 1994), and randomized experiments have even shown a clustering pattern emerges in suicidal ideation (Joiner, 2003).

Proposed mechanisms for suicide clustering are numerous. The most popular accounts involve some kind of *contagion* process, by which exposure to suicidal behavior transmits novel risk for suicide to those who are exposed (e.g., due to social learning, projective identification, priming, imitation, or some other mechanism; Haw et al., 2013). However, a more parsimonious account of suicide clustering involves simple *homophilic preference* (or more simply, *homophily*), a social phenomenon in which people prefer to associate with others who are similar to themselves (i.e., 'birds of a feather, flock together'; Newman, 2003). In this way, already-suicidal people may be socially clustered, but no meaningful risk is produced by their interaction with one another – if they die at similar times or places, it is because they already had similar risk profiles and backgrounds. Finally, an even more basic account implies suicide clustering could simply be an artifact of shared history. For example, economic downturn could simultaneously increase the suicide risk in a group of workers laid off from the same large factory. Even if they never actually knew one another, they may end their lives in an apparently patterned way. Along these lines, many examples of suicide clustering that appear to result from contagion, could also result from other, simpler processes (Joiner, 1999).

In actuality, all of these possibilities are specific hypotheses about the origin of a more general phenomenon known as *assortative mixing*, or just simply *assortativity* (Aral, Muchnik, & Sundararajan, 2009). Assortativity is the observation that any two people with similar attributes are more likely to be linked in some way than any two people with dissimilar attributes. This pattern could be produced by one person influencing friends toward similarity (contagion), it could be the result of already-similar people seeking each other out (homophily), or it could simply be the result of shared history in a group of people. Regardless of its origin, assortativity is nearly ubiquitous in human social networks. In fact, it is so common that it is argued to be one of the statistical signatures that distinguishes human social networks from their non-social counterparts (i.e., technological or microbiological networks; Newman & Park, 2003).

For instance, assortative mixing has been documented across numerous and diverse types of human behavior, ranging from smoking to obesity (Christakis & Fowler, 2007, 2008). Even measures of happiness have been shown to be assortative in both online and offline networks (Bliss, Kloumann, Harris, Danforth, & Dodds, 2012; Fowler & Christakis, 2009). Perhaps even more surprising than its breadth, assortativity also exhibits substantial reach in human networks. In each of the examples above, people were observed to be more similar than would be expected by chance, even when there were three degrees of separation between them (Christakis &

Fowler, 2013). That is, the friends of our friends' friends are more like us (e.g., in terms of happiness) than would be predicted by chance alone, despite the fact that we will likely never meet them.

Quantifying assortativity

Although assortativity was only recently defined mathematically (Newman, 2002), readers should note it is actually already quite familiar to them. In fact, it is closely related to the Pearson correlation coefficient, which is why we label our assortativity estimates with '*r*' throughout this investigation. Indeed, it is even bounded by the interval -1 to 1, with a value of 1 representing complete similarity across links (e.g., a marriage network of only gay couples), -1 representing complete dissimilarity across links (e.g., a marriage network of only straight couples), and a value of 0 representing random pairing of people on either side of a link (e.g., a marriage network with a high proportion of bisexual members).

For example, consider Figure 1, in which an example network of seven nodes (e.g., people), in which each node is given a score on some attribute of interest (e.g., sex, scored 0 or 1, light or dark). Imagine also that this network includes six total links among those seven members and we assign numeric ID numbers to label each link for clarity. If we create a table, where each link is a row and the attributes of the nodes at either end of that link are columns, we will get two columns for our attribute of interest (e.g., the sexes of the people on either side of a link). Note that any node with multiple links will appear in this table multiple times, once for each link to which it is adjacent. When the data are arranged in this way, the network assortativity of a given node attribute is simply Pearson correlation between those two attribute columns.[1] Thus, assortativity is the correlation of different people's attributes across the links that connect them.

Working with assortativity

Because contagion, homophily, and shared history all give rise to assortativity, they are mathematically confounded in most observational contexts and cannot be reliably distinguished (Shalizi & Thomas, 2011). This counter-intuitive finding is admittedly frustrating. However, in the case of suicide, remaining agnostic about the origin of assortativity of suicidal behavior may not be a disadvantage. While most research on the assortativity of suicidal behavior has focused on identifying the origins of that pattern (Haw et al., 2013), more basic (and unanswered) questions about the fundamental pattern of suicidal assortativity continue to have growing importance for prevention. For example, mobile apps are especially scalable (Sanderson, 2009) compared to traditional mental health treatment, and they have now been shown effective for the mitigation of suicidal and non-suicidal self-injury in multiple randomized control trials (Franklin, Fox, et al., 2016). Harnessing multiple aspects of social media for large-scale risk detection and prevention efforts has been the focus of increased attention in recent years (Franklin, 2016; Luxton, June, & Fairall, 2012; Robinson et al., 2016), for many of the same reasons.

Delivery of these new techniques can potentially be achieved through existing digital mechanisms (e.g., targeted advertising on social media sites), but to whom should they be delivered? Effective models of assortativity offer an important basis for answering this question. If assortative patterns of suicide risk can be characterized mathematically, then the identification of one at-risk individual will also constitute the probabilistic identification of many others as

[1] Note, assortativity coefficients are more general than Pearson coefficients and can be computed on unordered nominal variables as well (e.g., race). There, the relation to the Pearson correlation, above will not hold. In such cases, as well as with binary variables, the interval bounding the assortativity coefficient may contract to be narrower than -1.0 to 1.0, but will never be wider (Newman, 2002).

well. Phrased somewhat colorfully, if the pattern by which birds of a feather flock together is discovered, then it is only necessary to find one bird to infer the location of the whole flock – even if the etiology of that pattern remains enigmatic. Thus, establishing effective statistical models of assortativity for multiple kinds of suicidal behavior has important prevention potential, especially for the administration of emerging population-scale interventions and screening approaches.

Unresolved questions in suicide assortativity

Unfortunately, while assortative clustering of suicide fatalities have been documented in several studies (Haw et al., 2013), little is known about the degree to which assortativity is involved in other forms of suicidal behavior, including attempts, ideation or other gestures (e.g., online verbalizations). This is likely because fatalities are more discrete than other forms of non-lethal suicidal behavior (e.g., ideation), and they are more likely to produce contact with the medical and legal systems in which they occur. Thus, they are also more likely to be included in population-sized datasets capable of uncovering clustering (e.g., datasets with contextual variables like time and location of event). In contrast, suicidal ideation and attempts also generally require disclosure on the part of people who experience them, which has historically implied traditional survey methods for collecting data on such behavior. Unfortunately, these are hampered by a range of shortcomings, including non-disclosure due to stigma and retrospective biases. Collecting a dataset that is sufficiently large to study assortativity in this relatively uncommon behavior, along with the necessary precision about the contexts in which the behavior occurred to assess whether such behavior is assortative, is thus rarely achieved .

Moreover, when research on assortativity in non-fatal forms of suicidal behavior has been conducted, the designs have so far been unable to assess how far into the network clustering extends. For example, in one study, college roommates who chose to live together were found to have more strongly correlated suicidal ideation scores than roommates who were randomly assigned (Joiner, 2003). However, this study did not incorporate information on participants' broader social networks, instead collecting only individual dyads. Thus, it still unknown whether suicidal behavior is assortative beyond just one degree of separation, as has been shown for positive emotional states, which are assortative even up to three degrees (Bliss et al., 2012; Fowler & Christakis, 2009). As we discuss below, the ability to assess for both subtler behavior and to assess its reach is one of the key advantages of data collected from online networks. For example, many suicide attempts have been disclosed online that likely would not have been registered in a medical database (Wood, Shiffman, Leary, & Coppersmith, 2016).

When assortativity is expected, it has also been difficult to assess what the actual dimension of similarity is: are people clustered along suicide-specific characteristics or merely along mood states that are correlated with those characteristics (e.g., depression)? For example, in a study of over 3,000 adolescents from 46 high schools, Randall, Nickel, and Colman (2015) used logistic regression to show that exposure to a friend's suicide attempt increased a participant's risk of suicide attempt over the next year. They attempted to assess for 'assortative relating' by controlling for 50 potential confounds (e.g., depression, anxiety), finding that exposure to suicide was still an important predictor of subsequent attempts after inclusion of control variables. But the technique employed – logistic regression – doesn't actually evaluate assortativity in any of the confounds they attempted to control. Specifically, they would need to model the similarity between the attempter and the participant on these confounds to demonstrate or refute assortativity. Instead, only the overall level of each confound observed in the participant was controlled, not the similarity (i.e., assortativity) between the attempter and the participant on that confound. Thus, it remains an open question whether suicidal behavior is assortative at all. It may in fact be that another dimension of behavior, such as low mood, is the primary driver of these relationships and suicidal behavior is ancillary. For effective testing of the assortativity in suicidal behavior, more rigorous assortativity-specific control procedures must be implemented.

This deficit in our understanding of assortativity in non-fatal suicidal behavior should be especially concerning to researchers, as histories of non-fatal suicidal behavior are widely regarded as the most potent risk factors for subsequent death by suicide (Ribeiro et al., 2016). Indeed, the reader will quickly recognize how improbable it is that any well-established suicide clusters emerged *sui generis*, with members sharing no precursor risk factors prior to their deaths. Clearly more plausible is the expectation that when a cluster of suicide fatalities occurred, some clustering of non-fatal suicidal behavior (e.g., ideation, attempts) preceded it. Thus, for an effective prediction and prevention scheme, quantitative modeling of cluster-like phenomena in non-fatal suicidal behavior is a high priority.

Suicide and Twitter

Social media platforms of all kinds have offered an unprecedented volume, velocity, and variety of data to social scientists (for introductions, see Kern et al., 2016; Kosinski, Matz, Gosling, Popov, & Stillwell, 2015; Russell, 2013). Among these, the most consistently studied is likely Twitter (Twitter, 2016a), a ‘micro-blogging’ platform in which users broadcast 140-character posts directly to one another, or to the entire user-base of the platform simultaneously. Twitter’s scientific popularity is likely owed to how easily accessible it makes its data, the fact that most data collection schemes can be implemented for free, and the ease of data management (Russell, 2013). For example, Twitter requires limits all posts to 140-character units, which are more easily stored and analyzed than a corpus of longform blogposts from different servers, each with their own formatting and embedded media (e.g., video, images).

Like other social media platforms, Twitter has also been a boon to suicide researchers, who can now observe the behavior of individual people unobtrusively, collecting time-sensitive information that might not otherwise be shared due to stigma. For example, Wood et al. (2016) were able to analyze 125 users who publicly announced they had attempted suicide, finding that there were distinct signals in their pre-attempt posts that could have been used to predict their attempts. This is a relatively large sample, given the base rate of attempts (Drapeau & McIntosh, 2016), and it is possible that many of those participants would never have otherwise been studied, especially if their attempts were not severe enough to produce contact with the medical system. Elsewhere, strong correlations have now been documented between the use of suicide-related keywords on Twitter and state-specific age-adjusted suicide rates (Jashinsky et al., 2014); users posting suicide-related content are now understood to exhibit distinct linguistic profiles from those who do not (O'Dea et al., 2015); and unique temporal posting patterns have been detected for service members who died by suicide, relative to controls who died from other causes (Bryan et al., 2017). Taken together, these results provide compelling support for the value of the verbal content people post on social media platforms, especially on Twitter. In a range of designs, their verbal behavior has provided unique insight into suicidal behavior of various forms.

Unfortunately, while the scraping and analysis of data retrieved from social network platforms has exploded in suicide research, our understanding of the relationship between social networks and suicide has not. To our awareness, still no one has studied the actual social network contained in their social network data. Given that the structure of social the social landscape plays a significant role even in self-destructive behavior of non-social bacteria, we argue this is a critical deficit in our understanding of suicidal behavior and address it directly in this investigation.

The present investigation

Despite some methodological shortcomings, existing research implies two important questions about the assortativity of suicidal behavior. First, is any form of suicidal behavior assortative beyond one degree of social separation, as has been observed for many other emotional, physiological, and behavioral states (Christakis & Fowler, 2013)? Second, does that assortativity persist, even after another plausible source of assortativity (i.e., mood) – has been controlled? That is, holding the distribution of mood in the network constant, will the observed assortativity in suicide-related behavior still be higher than chance? Following previous research on assortative social behavior in humans, we predict suicide-related behavior online will be significantly more assortative than chance, up to three degrees of separation, but not further (Christakis & Fowler, 2013). We predict the same pattern will hold, even when accounting for the distribution of mood in the network. To address these two questions, the present investigation will utilize novel bootstrapping methods for the evaluation of assortative mixing in social networks (Christakis & Fowler, 2013; Newman, 2002), using data from a large online social media network, Twitter (Twitter, 2016a).

Methods

Methodological overview

This investigation involves several steps, a visual depiction of which are given in Figure 2. As shown by the large arrows in the top-left side of the figure, we began by collecting a random sample of real-time posting activity from Twitter – specifically, two non-sequential 28-

day periods. The second period was a replication sample for the first. Concurrently, Mechanical Turk (MTurk; Amazon, 2016) raters produced *suicide-relatedness* ratings for each of the 10,000 most common words in contemporary English. When both processes were complete, a machine scored each Twitter post according to whether it used words that were determined by MTurk raters to be highly related to suicide and whether it used words from a pre-existing list of ‘sad’ words, which were later used to infer users’ general moods (second hypothesis below). Note, these sad words had been previously rated and validated by a large group of MTurk raters for their relatedness to happiness and sadness, similar to the way suicide-relatedness was measured here (Dodds, Harris, Kloumann, Bliss, & Danforth, 2011).

After each post was scored by machine, a new group of MTurk raters then evaluated a random sub-sample of machine-scored posts, to ensure the validity of machine-scoring. Twitter users were then given total scores for both suicide-related verbalizations (SRV) and for sadness, which were simply the total numbers of posts that used words for the suicide-related and sad word lists, respectively. After user SRV and sadness scores had been totaled for each Twitter user, we estimated the social network. Specifically, we defined two users as linked if they engaged in at least one reciprocal reply with one another (i.e., Alice replies to at least one of Bob’s posts and Bob replies to at least one of Alice’s posts).

Analysis of the resulting network involved three phases. First, we estimated the observed assortativity (*r*) of SRV in the network using standard formulas (Newman, 2002), repeating the process for one to six degrees of separation (i.e., the approximate average distance between any two people in the world (Dodds, Muhamad, & Watts, 2003; Travers & Milgram, 1967). Second, we bootstrapped what the SRV assortativity distribution would look like if SRV posting were random in the network – the *marginal null distribution* – and assessed whether our observed *r* was significantly different from that distribution. Third, we bootstrapped what the SRV assortativity distribution would look like if SRV scores were conditional on sadness scores in the network – the *mood conditional null distribution* – and then re-assessed whether our observed *r* was significantly different than that distribution as well. Thus, the observed *r* at a given degree of separation remains constant, but is compared to two different null distribution benchmarks: (1) the *marginal* benchmark distribution and (2) the *mood conditional* benchmark distribution. This process was conducted at different degrees of separation, to assess whether SRV assortativity wanes with increasing social distance in the network. We evaluated one through six degrees of separation, as six is expected to be the approximate average social distance between any two people in the world (Dodds et al., 2003; Travers & Milgram, 1967).

Participants

The participants under investigation in this study were all active, English speaking users of Twitter. Twitter is a social media network of users who broadcast posts called ‘tweets’ (i.e., 140-character messages) to one another, or to the broader network of users (Twitter, 2016b). At the conclusion of 2015, it was estimated to have approximately 300 million monthly active users, worldwide (Twitter, 2016a). Through the Twitter Streaming Application Programming Interface (API), Twitter makes a random sample of the real-time activity inside the network available to anyone with an active Twitter account and a small amount of programming knowledge (Twitter, 2016c). The available information available through the API includes the text of posts, along with basic information about those posts (e.g., language, date created) and the users who produced them(e.g., number of friends, account age) that has not been explicitly flagged by the user as “private.” Importantly, research on other social media platforms indicates that although most users are familiar with available privacy settings, few users choose to engage them

(Debatim et al., 2009), suggesting that being unable to collect ‘private’ information on users would not likely have a substantial effect on data collection.

Through the Twitter Streaming API, we recorded all tweets flagged as having occurred in English for two non-sequential periods, 28 days each. The first period lasted from July 9, 2016 to August 6, 2016. The second period lasted from September 12, 2016 to October 10, 2016, and was used to verify the first period’s results were replicable. Previous research has shown that assortativity in mood on Twitter can be detected over time scales of days, weeks, and months (Bliss et al., 2012), suggesting our choice interval was appropriate for the hypotheses above.

This collection procedure yielded a final dataset of 64,499,981 posts, produced by 17,438,868 unique users. In the first period, there were 11,112,312 unique users; in the second period, there were 11,002,813 unique users; and 4,676,257 unique users were observed in both. Among all posts, 8,998,898 (14%) were marked by Twitter as replies to other users, 475,469 (<1%) were marked as coming from verified accounts (e.g., organizations, celebrities, other professional-level accounts), and 30,882,608 (48%) began with the letters *RT*, indicating they were retweets (i.e., re-posts) of other posts - these were not counted as replies for establishing links between users.[2] There were 456,358 user interactions that involved at least one reciprocal reply between users. This is the total number of links in the network per our definition. Approximately 20% percent of posts were produced through the Android app, 44% through the iPhone or iPad apps, 14% through the web app, and 21% were produced through an alternative source, like a computer visiting the Twitter website itself. The mean number of posts across all users was 3.70 (SD = 11.82) with a median of 1 post per user. Thus, this distribution had a strong positive skew, with 50% of users posting approximately one time, but with the top 5% posting

[2] Retweets were, however, counted toward each user’s overall sadness and SRV scores. This decision is addressed in more detail in the discussion.

more than 12 times throughout the entire study period. The mean and median account ages were 3.28 years (SD = 2.29) and 3.15, respectively.

Notably, other more traditional demographic information about the users in this sample (e.g., gender, age, race) is not available through the Twitter API. For that reason, the exact age, gender, and racial distributions of the users in this dataset are unknowable. However, previous research has indicated that, in general, Twitter users are more likely to be urban and male than the US general population (Mislove, Lehmann, Ahn, Onnela, & Rosenquist, 2011). It is also known that their racial and ethnic patterns differ from the typical geographic distribution observed in the US, with African American Twitter users over-represented in the Western US and users of Asian or Hispanic descent being over-represented in the eastern US (Mislove et al., 2011).

Raters

In addition to the Twitter users (participants) described above, this investigation also recruited two distinct groups of Mechanical Turk raters (MTurk; Amazon, 2016). The first (n = 2,429) was used to create a measure of whether posts included suicide-related words, which was used in the scoring procedure described below. Notably, word ratings from MTurk raters have previously been shown to be highly convergent with ratings from multiple alternative samples, when performing similar tasks to that employed here (see Dodds et al., 2011). The second (n = 162) was used as a validity check to evaluate whether the scoring procedure described below indeed tracked intuitive differences in perceived risk of suicidal behavior from posts' users.

Measures

Sadness and happiness. Previous research has established a measurement system – termed a *hedonometer* by its creators – for the assessment of mood in online networks, including

Twitter (Dodds et al., 2011). This hedonometer was constructed by first collecting over 10,000 of the most frequently used words in English, including digital ‘slang’ (e.g., *lol*, *:p*). These researchers then recruited a group of MTurk raters to rate each word on a scale of 1 *(sad)* to 9 (*happy)* several times, retaining each word’s mood average. The result is a kind of dictionary, where each word is associated with a mood score from 1 (*sad*) to 9 (*happy*). These mood averages were observed to be highly correlated (i.e., $r > .90$) with previous research on the emotional valence of common English words, as rated by university undergraduates (Bradley & Lang, 1999; Dodds et al., 2011), suggesting the mood quality of such words is consistent across groups.

Words with scores between 4 and 6 on the hedonometer (e.g., *manufacturers*, *diameter*) are typically treated as *stop words* (i.e., mood-neutral words) and removed before mood-based analysis because of their general irrelevance to either happiness or sadness (Bliss et al., 2012; Dodds et al., 2011). Functionally, this exclusion yields two lists: one of highly ‘sad’ words and one of highly ‘happy’ words. Here, we are primarily concerned with sad mood states. Thus, we regard a post as *sad* (scored 0, 1), if it includes any word with a mood score below 4 (n words = 1,043; 10% of the entire corpus). A user’s overall sadness score is thus simply the number of posts that were marked as sad during a given study period. All posts were scored in the same way, regardless of their type (e.g., original, reply, retweet). Note, we did not estimate a user’s overall happiness, as it is beyond the aims of this investigation; however, we do utilize happy words to help rule out sarcastic or colloquial talk about suicide-related content, as below.

Suicide-related verbalization (SRV) scores. The present investigation extends work of Dodds et al. (2011) on mood measurement to the assessment of suicide-related verbalizations in online social media. This process consisted of a measure creation phase and a validity check phase. First, the same 10,000 words utilized for the hedonometer above were re-rated by a new group of MTurk raters (n = 2,429) for their relatedness to suicide. Each rater responded to approximately 100 randomly selected words, for which they were asked: *how closely related are each of the following words to "suicide"?* Ratings were made on a 1 (*not at all related*) to 9 (*highly related*) scale and the mean rating for a given word was retained. The average reported age of raters was 36.43 (SD = 12.22), 56% reported being female, and 82% reported being white, 7% reported being Hispanic or Latino. Raters were paid $0.50 for their ratings.

Following previous studies using the hedonometer (Bliss et al., 2012; Dodds et al., 2011), we define stop words that are excluded from analyses. Specifically, we remove any word with an average suicide-relatedness score at or below 5 as negligibly related to suicide, retaining only words with a score greater than 5. Note, we utilize a different cut point than previous research (Dodds et al., 2011), given our scale is unipolar (not suicide-related to highly suicide-related) rather than bipolar (sadness to happiness) and because 5 represents the point on our scale where is word is closer to *highly related to suicide* than *not at all related to suicide*. When scoring SRV for each post (0 or 1), we further ruled out posts that included any happy word (i.e., any word with a hedonometer mood score above 6). This latter standard was included to help exclude disingenuous or otherwise colloquial suicide-related posts (e.g., *I'm so happy I could kill myself*).[3] Thus, a post was treated as SRV-positive (scored 1) by our machine-scored algorithm if it (a) contained at least one word with a suicide-relatedness score greater than 5 and (b) did not contain any happy words (with hedonometer scores above 6), outlined above. It was regarded as SRV-negative (scored 0), otherwise. Examples of hypothetical posts, how they would be scored,

[3] Note, no additional criteria of this kind were applied to the scoring of sad posts. This was to maintain comparability to previous research using the hedonometer, in which post mood scores were assigned based only one which words are present in a post (regardless of which words are absent).

and how those scores are aggregated to create user total scores is depicted in Table 1. All scoring was conducted by machine, using an R script run on an Elastic Cloud Computing (EC2) server hosted by Amazon Web Services (Amazon, 2017; R Core Team, 2017). Again, all posts were scored in the same way, regardless of their type (e.g., original, reply, retweet).

***SRV descriptive statistics*.** Following the lexicon creation procedure described above, a sample of MTurk raters (n = 2,429 rated the 10,222 most common words in English for their relatedness to suicide. Each word was rated an average of 20.61 times (SD = 4.52), with a minimum of 5 ratings for any word overall. Ninety-eight percent of words received an average rating less than or equal to a 5 on the 1 to 9 scale. The remaining 219 words with an average rating greater than 5 were denoted as words with high suicide-relatedness (see Figure 3).

In both study periods combined, the SRV post scoring algorithm described above flagged 6,453,415 (1%) posts as including at least one word with high suicide-relatedness and 24,072,770 (37%) posts as including at least one sad word. There were 54,750,738 (85%) posts marked as including a happy word. These findings are consistent with previous research showing English is a positively valenced language (Dodds et al., 2011), as well as research suggesting the short-term prevalence of suicide-related behavior like suicidal ideation is relatively uncommon (Nock et al., 2008) and research showing words specific to suicide are relatively uncommon among languages of European origin (Daube, 1972). Among posts with at least one highly suicide-related word, 5,638,419 (87%) also included a happy word and were ruled out, resulting in a total of 814,996 (1%) of posts marked SRV-positive. Conversely, 742,308 (91%) of SRV-positive posts also included at least one sad word. SRV-positive posts were produced by 688,610 (<1%) unique users throughout the entire study period. Posts that scored positive for SRV came from accounts that were slightly older (d = 0.06), but had a similar number of total posts (d <

0.01), a similar number of people followed (*d* = 0.02), and a similar number of followers (*d* < 0.01), compared to SRV-negative posts. SRV-positive posts were somewhat more likely to be produced with Apple products (iPhone or iPad = 52%) than SRV-negative posts (iPhone or iPad = 44%).

SRV validity. Our scoring system was designed to aid in an initial proof of concept: that verbal behavior related to suicide is assortative in large social networks. It was not designed to be a risk assessment measure with diagnostic-grade precision. However, we did perform a range of validity checks to ensure the instrument was measuring a suicide-relevant behavior with potential clinical importance.

First, if using words with high suicide-relatedness roughly tracks a construct analogous to suicidal ideation, then it should have a moderate (but not high) correlation with hedonometer mood scores among sad words (convergent validity); it should also have negligible correlation with hedonometer mood scores among happy words (divergent validity). This latter expectation arises from the assumption that the move from 'depressed' to 'disappointed' likely impacts suicidal behavior of all forms more than the move from 'happy' to 'joyful'. As shown in Figure 3, this is indeed what we find. Among only sad words (i.e., mood score below 4), the absolute Spearman correlation between a word's suicide-relatedness and mood score was ρ = .38 (*p* < .001); among only happy words (i.e., words with a mood score above 6), that correlation was ρ = -.06 (*p* < .001). Figure 3 also refutes the concern that 'all suicidal words are already sad words,' showing that there are words that fall in the high suicide-relatedness range and not the sad range, as well as the reverse. Thus, the set of words the individual word-ratings for our SRV measure demonstrate both initial convergent validity (i.e., correlated in the appropriate context) and initial divergent validity (i.e., uncorrelated in the appropriate context) with an existing measure. Figure

3 demonstrates this visually, in which existing hedonometer scores are shown to be related to suicide-relatedness scores only at the low (sad) end of the spectrum.

The second validity check directly evaluated whether there was differential information in SRV-positive and SRV-negative posts themselves. That is, after scoring each post as SRV-positive or negative by machine, we asked a new group of SRV-blind MTurk raters (n = 162) to evaluate a random subsample of these posts (n = 1500, 50% SRV-positive). Specifically, raters were asked “is the author of the following post at risk for suicide?” and were instructed that “posts that are sarcastic, informational, or totally unrelated to suicide should be rated lower; posts that appear to imply risk for suicide should be rated higher.” Ratings were made on a 0 (*extremely unlikely*) to 5 (*extremely unlikely*) scale. Raters were paid $0.50 for their ratings. Each rater rated approximated 65 posts. The average reported age of raters was 35.96 (SD = 12.39), 66% reported being female, and 80% reported being white, 3% reported being Hispanic or Latino.

If scoring a post SRV-positive tracks something like suicidal ideation or suicide risk more broadly, then the users who authored SRV-positive posts should be perceived by external observers of their posts to be at greater risk for suicide than users who authored SRV-negative posts. Indeed, that is again what we found. Results show SRV-positive posts are perceived by MTurk raters (n = 162) as having been produced by users with significantly higher risk for suicide (*b* = 0.59, SE = .04, *p* <. 001, Cohen’s *d* = .82), suggesting our scoring rule broadly tracks the kind of intuitive judgements about suicide-related behavior that would be made by humans who are hand-rating social media-posts. This indicates at least two things. First, SRV-positive posts are unlikely to be dominated simply by casual or sarcastic comments, with raters having been explicitly told to rate such comments as having lower risk. Second, SRV posts – at least on-face – involve more concerning suicide risk than non-SRV posts making them a worthwhile proxy for this initial investigation on assortativity in suicide-related verbal behavior. raters

Network assembly

Data streamed from Twitter include both the text of individual posts and information about those posts, such as the user account that produced them. While most posts are general broadcasts to the entire network, some posts are designated by the user as replies to the posts of other users (e.g., Alice posted X, but Bob replied Y to Alice's X post; (Twitter, 2016b). Following Bliss et al.'s (2012) work on the assortativity of mood on Twitter, we designated two users as *linked* in a network when they have each replied to one another at least once. For example, Alice and Bob are linked if (1) Alice has replied to at least one of Bob's posts and (2) Bob has replied to at least one of Alice's posts (i.e., reciprocal-replies), in the period under consideration. This standard thus requires that any users regarded as linked must have deliberately interacted with one another at least once.

Network analysis

Observed assortativity. A network is simply a collection of nodes (people) and the links (interactions) among them (Newman, 2010). Here, the nodes also have attributes of interest: their total numbers of SRV-positive posts and sad posts. The propensity of linked nodes to have similar numbers of SRV-positive posts is estimated with the assortativity coefficient, *r*, detailed in the introduction (Newman, 2002).

Degrees of separation. Assortativity quantifies the similarity of nodes across links in a network. To assess how that quantity changes at increasing levels of social distance, we estimate the assortativity coefficient for each network at 1 through 6 degrees of separation. That is, we calculate the assortativity of direct ‘friends’ (links of degree 1), ‘friends of friends’ (links of degree 2), ‘friends of friends of friends’ (links of degree 3), and so on up to degree 6, which is thought to be the most plausible upper limit for the average social distance between any two people in the world (Dodds et al., 2003; Travers & Milgram, 1967). This is a conservative upper bound and the actual upper limit of social distance in online networks may be as low as three or four degrees (i.e., it may be possible that the true average degree of separation between people online is as low as three; Backstrom, Boldi, Rosa, Ugander, & Vigna, 2012). For clarity, we denote r_d as the assortativity (*r*) of SRV in a network, at *d* degrees of separation (e.g., ‘r_3 = .15’ would signify that the assortativity of SRV at three degrees of separation is .15)

Hypothesis tests. Once the assortativity is estimated for a given observed network (e.g., from the first study period) at a given degree of social distance (e.g., 2), it is important to assess whether that observed level of assortativity could plausibly be the result of chance. The observed assortativity coefficient was thus compared to two benchmarks. The first benchmark was the *null marginal assortativity distribution* (or simply, the marginal null), which represents the range of values an observed network’s assortativity coefficient could take on if (a) the network structure remained the same, but (b) SRV scores were assigned to the nodes at random. Following previous research, this is achieved by bootstrapping (Christakis & Fowler, 2013). Specifically, we take the observed network and hold the links among people constant, but shuffle their SRV scores many times (with replacement). By recording what the assortativity coefficient would have been during each shuffle, we achieve a null distribution of what ‘random’ assortativity would look like and can estimate a bootstrapped *p*-value for the observed assortativity coefficient. Specifically, if our observed assortativity coefficient was in the top 2.5% of the null marginal distribution, we regarded it as significantly different from the marginal null.

The second benchmark was termed the *null conditional assortativity distribution* (or simply, the conditional null), which represents the range of values an observed network's assortativity coefficient could take on if (a) the network structure remained the same, (b) the user sadness scores in the network remained the same, and (c) the SRV scores were shuffled within sadness score groups. This is the null distribution of SRV, conditional on mood. Phrased statistically, the conditional null represents what SRV assortativity scores would look like if SRV were random, once the sadness scores of users were controlled. Phrased practically, the conditional null is the distribution of what SRV assortativity scores would look like if users SRV-positive posting remained correlated with their own sadness scores, but were not at all associated with the SRV-related posting of any other users. Thus, by shuffling SRV scores within each sadness level and recording what the assortativity coefficient would have been during each shuffle, we achieve a null distribution of what mood-conditional random assortativity would look like. Consistent with standard hypothesis testing, an observed SRV assortativity value is regarded as significantly different from a null distribution (i.e., the totally random marginal null or the mood conditional null), if it occupies a low-probability region of that distribution – either near the end of or completely off the tail (i.e., it is as extreme or more than the top 2.5% of bootstrapped samples). Thus, if our observed assortativity coefficient was in the top 2.5% of this null conditional distribution, we regarded it as significantly different from the conditional null. This is analogous to concluding the assortativity of SRV is even stronger than would be expected, given the distribution of mood scores in the network.

Results

Results of the assortativity hypothesis tests are depicted in Figure 4, which displays the plausible range of assortativity values (horizontal axis) for the marginal null distribution (dark) and the conditional null distribution (light), along with the observed SRV assortativity coefficients (points) for each degree of separation (vertical axis). Again, an observed SRV assortativity value is regarded as significantly different from a given null distribution, if it occupies a low-probability region of that distribution – either near the end of or completely off the tail.

Observed assortative distance

As expected from previous research (Bliss et al., 2012; Christakis & Fowler, 2013), observed SRV assortativity values generally decline in magnitude at each increasing level of social distance (degrees of separation).[4] Users' SRV scores are most like those with whom they interact directly (r_1 = .08 - .10), slightly less like the SRV scores of those who are two degrees away (r_2 = .06), and least similar to the scores of those who are six degrees away (r_6 = .01 - .02). Of note, because the SRV scoring algorithm utilized in this investigation was a relatively course-grained (i.e., based on the presence and absence of key words), it likely includes notable random error variance, which is known to shrink measures of association like correlations (Casella & Berger, 2001), the family to which assortativity coefficients belong (see above). As such, there is good reason to expect the assortativity coefficients observed in this investigation are conservative estimates of the true assortativity of suicide-related behavior on Twitter or in other online and offline networks. Future measures with more diagnostic-grade precision are likely to produce even higher assortativity values.

Observed assortativity versus marginal null (random SRV)

[4] Note, a few small wrinkles are observed in this pattern. For example, the transition from three to four to five degrees of separation in Period A is not a perfectly decreasing trend (i.e., four degrees is slightly less assortative than five). However, statistical noise is likely the best explanation for this small deviation from the broader expected pattern, given that unexpected trend was not observed in both periods.

Recall, the first hypothesis in this investigation was that SRV would be significantly assortative, up through three degrees of separation. Results show this prediction was met and exceeded, with SRV demonstrating significant assortativity all the way through six degrees of separation. Specifically, all observed assortativity values were as high or higher than the top 2.5% of bootstrapped values from the null marginal distribution, regardless of study period or degree of separation. Statistically, this indicates socially linked people are more similar on SRV (e.g., low with low, high with high) than would be expected by holding the network constant and randomizing individual SRV scores; that is, the SRV assortativity on Twitter is significantly greater than random chance alone and that this effect remains significant at least up to six degrees of separation.

Phrased practically, a randomly selected person from this network is more similar to the friends of their friends' friends' friends' friends' friends than would be expected by chance. From a prevention perspective, this implies that there is information on that person's suicide-related behavior that we can infer from the behavior of people they have never known and never will. We further demonstrate the significance of this finding for prevention with an illustrative simulation example later in the Implications section of the Discussion.

Observed assortativity versus mood-conditional null

The second hypothesis in this investigation was that SRV would remain significantly assortative, even after mood was controlled in the network. In both study periods, observed assortativity values were significantly greater than would be expected at both 1 and 2 degrees of separation, even when sadness scores were controlled. In other words, the assortativity of SRV at 1 and 2 degrees of separation is too high to be a plausible artifact of mood assortativity. Thus, results indicate the observed assortativity of SRV on Twitter is not statistically attributable to the already well-documented assortativity in mood (Bliss et al., 2012).

However, at social distances greater than 2, the observed assortativity at each increasing degree of separation was typically significant in only one of the two study periods, with only degree 5 being significant in both periods. Absent a compelling justification for why 5 degrees of separation are in some way unique from 3, 4 and 6 degrees, a scientifically conservative interpretation of these results is that the observed assortativity coefficients for degrees 3 through 6 cannot be reliably distinguished from what would be expected given the already established assortativity of mood. That is, following the often neglected insight that negative scientific evidence is more powerful than positive (Meehl, 1978), we interpret these mixed results for degrees of 3 and above as functionally null.

Taken together, results are consistent with the hypothesis that SRV is significantly more assortative than pure chance (marginal null), all the way up to 6 degrees of separation. In addition, the observed assortativity in SRV is also significantly greater than would be expected based on the known assortativity in mood and the expectation that low mood states are associated with various other forms of suicidal behavior (conditional null). However, SRV is only more assortative than would be expected from mood, for 1 and 2 degrees of separation. For 3 or more degrees of separation, SRV is still assortative, but could plausibly be an artifact of social connections driven by mood alone.

Discussion

This study investigated the connection between a social network (i.e., the Twitter network) and the suicide-related behavior enacted by individuals who occupy it (i.e., verbal behavior involving suicide-related words; SRV). Although several studies have previously attempted to evaluate the similarities of people at risk for suicide or the clustering of suicidal behavior, to our awareness this is the first study to directly estimate the assortativity of any kind of suicide-related behavior. It is also among the first of its kind to evaluate suicide-related behavior in a network of this scale, having analyzed over 64 million posts from more than 17 million unique users in two 28-day periods. Consistent with the expectations of previous research on the assortativity of mood (Bliss et al., 2012; Fowler & Christakis, 2009), SRV was observed to be significantly more assortative than would be expected by chance alone – remaining significant up to at least six degrees of separation between individuals in the network. When mood was controlled, the assortativity of SRV was still significantly higher than would be expected by chance, up to 2 at least degrees of separation.

Taken together, our results suggest that verbal behavior related to suicide (i.e., SRV) occurs in a socially assortative patterned way on Twitter, with individuals who engage in such verbal behavior having a significantly higher likelihood of interacting with one another than chance. Although the magnitude of this correlation-style estimate was modest (i.e., r_1 = .07 - .10), its reach was quite substantial – observed at up to 6 degrees of separation. Moreover, in the neighborhood encompassed by two degrees of separation, the assortativity of SRV was not plausibly be attributable to the already known assortativity of mood. The remainder of this discussion addresses the implications of these findings, along with caveats to their interpretation and directions for future research.

Implications

There are two primary implications arising from the findings presented earlier: one pragmatic and one etiological. Pragmatically, the results imply that for at least one form of suicide-related behavior, ‘birds of a feather flock together’ in a statistically regular way.

Assuming more serious forms of suicidal behavior (e.g., suicide attempts) follow a similar pattern – a priority question for future research - only a small number of ‘known birds’ are needed to locate much of the ‘flock.’ We emphasize that this is true, regardless of the causal origin of that assortative ‘flocking’ pattern.

The most obvious way to exploit such a pattern is to use snowball sampling when screening for suicide risk. To see this, consider Figure 5, which depicts a random subset of the first period’s network. In this subsample, there are 107 total people, 29 of which produced at least one SRV-positive post during the study period (dark nodes). Among those 29 SRV-positive users, 17 knew at least one other SRV-positive user (highlighted in red). Thus, the probability an SRV-positive user knew at least on other is 17/29 = .59 and the probability an SRV-negative user knew at least one SRV-positive user is 42/78 = .54. This difference may appear numerically modest, but it is practically quite powerful.

For example, imagine the network in Figure 5 represents a small high school, in which one student has recently taken his life. Concerned about the suicide risk of the remaining students, an administrator asks the district’s lone school psychologist to locate any other students are at risk so they can be referred for psychotherapy. Unfortunately, the school psychologist knows she only has the time and resources to perform 25 total screens, guaranteeing 82 students will remain unscreened. She has a list of every student in the school, but how does she choose who to screen?

The most common approach is likely simple random screening (without replacement), in which the school psychologist chooses 25 people from her list in advance, then screens them one after another. Such a scenario will follow a hypergeometric distribution (Casella & Berger, 2001), where the mean number of true positives she will locate after 25 screens is 6.78

(sensitivity = .23). However, if she is aware that suicide risk has non-zero assortativity, she can exploit that pattern with snowball sampling: (1) ask a teacher who the decedent's closest friends were and screen them, (2) ask any positives from that group to list their closest friends and screen those friends, (3) ask any positives from that new group who their closest friends are, screen them, and so on, (4) if ever there are no more positives in a friend group, screen students at random until a positive is found and begin the snowball procedure again until resources run out.

If this strategy is simulated 1,000 times using only the data in Figure 5, the average number of true positives using 25 screens is 8.30 (sensitivity = .29), about 22% higher than random sampling's average. Beyond the averages, comparing their overall distributions shows that exploiting even the modest assortativity of $r_1 = .12$ from this network will cause the snowball-style screening to find more true positives than random sampling 68% of the time, even though the same total number of people are screened in each case.

Readers will also note these results are from a small network, with an even smaller number of available screens. But as the size of the network and available screens grows (e.g., online) the superiority of the assortativity-informed snowball approach accelerates non-linearly. For example, if there were 1,000 screens applied to the whole first-period network (n = 168,098, 21% SRV-positive), the random sampling approach would find on average 207 true positives, but the snowball approach would find and average of 358 (across 10 simulated runs), a 73% improvement.

Aside from the obvious benefits of enhancing the proportion of true positives located during mass screening, capitalizing on assortativity is likely to have other benefits for suicide research as well. For example, researchers struggling to acquire participants from a low-base rate group like suicide attempters could employ assortativity-informed recruitment strategy like the snowball approach above.[5] Second, and more speculatively, it is possible that assortativity-informed interventions could produce positive ripple effects throughout the social network after the treatment of only a few at-risk nodes. Such a possibility would improve both issues related to limited resources and difficulties reaching at-risk people who wouldn't otherwise present for treatment (i.e., because they would indirectly receive a dose of treatment through their friends who were treated). We emphasize, however, that this treatment approach first requires confirmation about the causes of suicide-related assortativity in social networks, which the current study was not designed to disentangle. While the current results provide an initial hint of an assortativity-informed treatment approach, additional research will be required prior even to pilot interventions of this kind.

A final comment here, although our school example above was employed only for its simplicity, the reader may still wonder: what do the current results say about the assortativity of suicide-related behavior in offline networks? In the interest of transparency, we report the current results say little about behavior in offline networks because no offline behavior was studied here. However, as we explained in the introduction, assortativity is one of the defining statistical signatures of a vast majority of human social networks studied so far (Newman, 2002, 2003). Thus, there is good reason to expect that if it has been observed online, it will be observed offline as well. For example, previous research on mood has shown it to be assortative in both online and offline contexts (Bliss et al., 2012; Fowler & Christakis, 2009). In addition, the observed

[5] Researchers concerned about this admittedly non-random sampling approach are reminded that (a) their existing samples were already highly likely to be non-random, (b) this kind of non-random sampling bias is better than the status quo because it is quantifiable and thus able to be accounted in the final model (Bareinboim & Pearl, 2012), and (c) knowledge of the social relationships between participants in a study actually opens a whole new dimension of investigation (e.g., the effects of the network on the study topic).

clustering of suicides in both time and space imply that at least some kind of assortative pattern must apply to at least some kinds of suicidal behavior (Haw et al., 2013).

Although further research is needed, there is good reason to expect some forms of suicidal behavior (e.g., ideation, attempt risk) in offline networks to be as assortative as we observed here, if not more so. Indeed, we even suspect that the long-form interaction available offline will facilitate even higher assortativity than we observed on Twitter, where interactions are limited only to 140-character bursts available for all the public to see. Instead, the privacy, information density, and potential for a much higher number of interactions (albeit with fewer others) in offline social relationships likely allows for much more assortativity in suicidal behavior to emerge over time and we look forward to future research investigating this possibility.

The second implication of these results is related to our more basic understanding of the etiology of suicide. Specifically, the mood-controlled results suggest that the observed clustering of suicide-related behavior in the social environment is not merely a side-effect of pre-existing clusters of people with similar moods, who also happen to be at greater risk for suicidal behavior because their moods are low. Rather, results imply that a suicide-specific mechanism is likely to be involved in the assortative clustering of at least suicide-related verbal behavior.

Thus, while it is not currently possible to disentangle the causal mechanism of assortative patterns in non-experimental network settings (Shalizi & Thomas, 2011), the current results have still narrowed the search. That is, whether the cause of this assortative pattern is social contagion, homophily, or shared history, it is likely to be a suicide-specific class of one of those mechanisms. For example, people may exhibit similar SRV scores because one high-SRV member of the network unintentionally influences his neighbors to engage in such suicide-related verbal behavior as well (social contagion); however, the present results imply it is unlikely that a the similarity of SRV scores among socially connected people is merely the result of one having influenced the others to become sad, which had the secondary consequence of also increasing their propensity to utilize highly suicide-related words.

Interpretive caveats and future directions

As described above, this study breaks ground in new area of suicide research, utilizing a model capable of reconciling large group-level patterns with the individual-level behavior that gives rise to them; however, as an initial study in this area, several caveats are important for the interpretation of results. Most importantly, we employed a relatively coarse-grained measure of individual behavior, SRV. This was intentional, as the present study was designed to demonstrate the significance of potential large-scale assortativity in suicidal behavior, rather than to screen and refer people for intervention. However, the limitations of the present study do still imply a few important agenda items for future research to resolve.

First, it will be important to confirm that more refined methods for measuring suicide-related verbal behavior online remain assortative, like the SRV measure here. For example, an intuitive next step would be to assign each post a score (e.g., cosine similarity) based on its grammatical similarity to a known set of suicide notes (Pestian et al., 2012). Using such scores, it would be possible to ask whether people with a propensity to write content with high similarity to a suicide note associate more often with one another than chance. More directly, is suicide-note-like posting also assortative?

A related question is what kinds of posts should be included in any kind of SRV calculation? For example, this investigation counted ‘retweets’ – cases where one user re-posted another user’s original post – in the sadness and SRV calculations. Thus, if one user produced an

SRV-positive post and a second user retweeted that post, the second user's SRV score would increase. This was done to treat all posts as equal, as little empirical means were available to prioritize one kind of post over another. We note this decision could be either desirable or undesirable for the measurement of SRV. On one side, being at elevated risk for suicide may increase the likelihood a user retweets a suicide-related post, making it important to include such re-posts in their SRV score. On the other side, it may also be that users predominantly retweet other users' suicide-related posts out of concern for those users, making such re-posting unreflective of the re-posters actual suicide risk. We concede this is an open question in need of empirical resolution. Depending on its resolution, the assortativity coefficients in this study could increase or decrease; however, it is our general expectation that they would increase, as resolving this question would reduce another form of random error variance, resulting in more precise measurement.

Second, it is critical to verify whether the kinds of suicide-related verbal behavior that are assortative are also predictive of more serious suicidal behavior offline. Specifically, there is already good reason to believe that online verbal behavior is predictive of offline suicide risk (e.g., Bryan et al., 2017; Wood, Shiffman, Leary, & Coppersmith, 2016), but it is not clear whether the features of verbal behavior that are risk-predictive are themselves assortative. For example, imagine the ratio of verbs to nouns a user posted online predicted offline suicide attempts in some way, but that verb-noun ratios were not assortative among users – there were no 'flocking' pattern. In such a case, the practical advantages accrued from network models would likely be null: assortativity-driven snowball screening would be no better than random screening, and would probably be worse.

Importantly, our validity check shows that external raters perceive the users who produced SRV-positive posts to be at greater risk for suicide attempts. Coupled with the fact that we observed reliable assortativity in those SRV scores across two independent periods, this implies that at least some features of suicide risk are likely to be assortative. However, a study that can directly test the assortativity of a strong index of suicide risk – or of suicidal behavior, directly – is an important next step in this line of research, without which strong recommendations cannot be made.

Lastly, because of the sheer volume of data, we were only able to evaluate one online network, Twitter. There are important features of Twitter that make it different from offline networks (e.g., the ability to directly approach and interact with celebrities) and from other online networks (e.g., nearly everything is public, posts are all sharply limited to 140 characters). Additionally, the demographic characteristics of at least US Twitter users are different from the general US population, a difference that may hold for other online and offline networks as well. For these reasons, a variety of network-related phenomena could likely manifest differently in other networks. Establishing quantitative models of their potential differences, is thus, a priority before augmenting preventions schemes in those networks with the approaches discussed here.

Conclusion

Assortativity is a hallmark of human social networks (Newman, 2002), observed in numerous and disparate forms of behavior (Christakis & Fowler, 2013). This investigation demonstrates that suicide-related verbal behavior is also among them, and suggests that other more serious forms of suicidal behavior might be as well. Notably, the coefficients in this investigation were of only modest magnitude, but they demonstrated substantial social reach –

lasting up to at least 6 degrees of separation. If exploited, such a pattern has the capacity to substantially improve screening approaches in both small and large networks alike, increasing the number of people at risk for a suicide attempt found in advance, using only existing resources.

References

Amazon. (2016). Amazon Mechanical Turk. Retrieved July 5, 2016, from https://www.mturk.com/mturk/help?helpPage=overview

Amazon. (2017). Amazon Web Services (AWS) - Cloud Computing Services. Retrieved September 18, 2017, from //aws.amazon.com/

Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, *106*(51), 21544–21549. https://doi.org/10.1073/pnas.0908800106

Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2012). Four degrees of separation. In *Proceedings of the 4th Annual ACM Web Science Conference* (pp. 33–42). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=2380723

Bareinboim, E., & Pearl, J. (2012). Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics* (pp. 100–108). Retrieved from http://www.jmlr.org/proceedings/papers/v22/bareinboim12/bareinboim12.pdf

Bliss, C. A., Kloumann, I. M., Harris, K. D., Danforth, C. M., & Dodds, P. S. (2012). Twitter reciprocal reply networks exhibit assortativity with respect to happiness. *Journal of Computational Science*, *3*(5), 388–397. https://doi.org/10.1016/j.jocs.2012.05.001

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical report C-1, the center for research in psychophysiology, University of Florida. Retrieved from http://api.rue89.nouvelobs.com/sites/news/files/assets/document/2010/07/bradley1999a_1.pdf

Brent, D. A., Kerr, M. M., Goldstein, C., Bozigar, J., Wartella, M., & Allan, M. J. (1989). An outbreak of suicide and suicidal behavior in a high school. *Journal of the American Academy of Child & Adolescent Psychiatry*, *28*(6), 918–924.

Bryan, C. J., Butner, J. E., Sinclair, S., Bryan, A. B. O., Hesse, C. M., & Rose, A. E. (2017). Predictors of Emerging Suicide Death Among Military Personnel on Social Media Networks. *Suicide and Life-Threatening Behavior*. https://doi.org/10.1111/sltb.12370

Casella, G., & Berger, R. L. (2001). *Statistical Inference* (2nd edition). Australia ; Pacific Grove, CA: Duxbury Press.

Centers for Disease Control and Prevention (CDC). (2016). Web-based Injury Statistics Query and Reporting System (WISQARS). Retrieved January 1, 2016, from http://www.cdc.gov/ncipc/wisqars

Christakis, N. A., & Fowler, J. H. (2007). The Spread of Obesity in a Large Social Network over 32 Years. *New England Journal of Medicine*, *357*(4), 370–379. https://doi.org/10.1056/NEJMsa066082

Christakis, N. A., & Fowler, J. H. (2008). The Collective Dynamics of Smoking in a Large Social Network. *New England Journal of Medicine*, *358*(21), 2249–2258. https://doi.org/10.1056/NEJMsa0706154

Christakis, N. A., & Fowler, J. H. (2013). Social contagion theory: examining dynamic social networks and human behavior. *Statistics in Medicine*, *32*(4), 556–577. https://doi.org/10.1002/sim.5408

Christensen, H., Cuijpers, P., & Reynolds, C. F. (2016). Changing the direction of suicide prevention research: A necessity for true population impact. *JAMA Psychiatry*. https://doi.org/10.1001/jamapsychiatry.2016.0001

Daube, D. (1972). The linguistics of suicide. *Philosophy & Public Affairs*, 387–437.

Deisenhammer, E. A., Ing, C. M., Strauss, R., Kemmler, G., Hinterhuber, H., & Weiss, E. (2009). The duration of the suicidal process: how much time is left for intervention between consideration and accomplishment of a suicide attempt? *The Journal of Clinical Psychiatry*, *70*(1), 19–24.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter. *PLoS ONE*, *6*(12), 1–26. https://doi.org/10.1371/journal.pone.0026752

Dodds, P. S., Muhamad, R., & Watts, D. J. (2003). An experimental study of search in global social networks. *Science*, *301*(5634), 827–829.

Drapeau, C. W., & McIntosh, J. L. (2015). *U.S.A. suicide 2014: Official final data*. Washington, DC: American Association of Suicidology. Retrieved from http://www.suicidology.org/Portals/14/docs/Resources/FactSheets/2014/2014datapgsv1b.pdf

Drapeau, C. W., & McIntosh, J. L. (2016). *U.S.A. suicide 2015: Official final data*. Washington, DC: American Association of Suicidology. Retrieved from http://www.suicidology.org/Portals/14/docs/Resources/FactSheets/2015/2015datapgsv1.pdf?ver=2017-01-02-220151-870

Farmer, R., O'Donnell, I., & Tranah, T. (1991). Suicide on the London Underground System. *International Journal of Epidemiology*, *20*(3), 707–711. https://doi.org/10.1093/ije/20.3.707

Fowler, J. H., & Christakis, N. A. (2009). Dynamic spread of happiness in a large social network: longitudinal analysis of the Framingham Heart Study social network. *BMJ: British Medical Journal (Overseas & Retired Doctors Edition)*, *338*(7685), 23–27.

Franklin, J. C. (2016). *How Technology Can Help Us Move Toward Large-Scale Suicide Risk Detection and Prevention*. Recorded Webinar presented at the American Association of Suicidology. Retrieved from http://www.suicidology.org/store/BKctl/ViewDetails/SKU/AASTECHRISKWE

Franklin, J. C., Fox, K. R., Franklin, C. R., Kleiman, E. M., Ribeiro, J. D., Jaroszewski, A. C., … Nock, M. K. (2016). A brief mobile app reduces nonsuicidal and suicidal self-injury: Evidence from three randomized controlled trials. *Journal of Consulting and Clinical Psychology*, *84*(6), 544–557. https://doi.org/10.1037/ccp0000093

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., … Nock, M. K. (2016). Risk Factors for Suicidal Thoughts and Behaviors: A Meta-Analysis of 50 Years of Research. Retrieved from http://psycnet.apa.org/psycinfo/2016-54856-001/

Goldsmith, S. K., Pellmar, T. C., Kleiman, A. M., & Bunney, W. E. (2002). *Reducing suicide: A national imperative*. Washington DC: National Academies Press.

Gould, M. S. (1990). Suicide clusters and media exposure. In S. J. Blumenthal & D. J. Kupfer (Eds.), *Suicide over the life cycle: Risk factors, assessment, and treatment of suicidal patients* (pp. 517–532). Arlington, VA, US: American Psychiatric Association.

Gould, M. S., Petrie, K., Kleinman, M. H., & Wallenstein, S. (1994). Clustering of attempted suicide: New Zealand national data. *International Journal of Epidemiology*, *23*(6), 1185–1189.

Gould, M. S., Wallenstein, S., & Davidson, L. (1989). Suicide clusters: A critical review. *Suicide and Life-Threatening Behavior*, *19*(1), 17–29.

Hannak, A., Anderson, E., Barrett, L. F., Lehmann, S., Mislove, A., & Riedewald, M. (2012). Tweetin'in the Rain: Exploring Societal-Scale Effects of Weather on Mood. In *ICWSM*. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewFile/4648/5036

Haw, C., Hawton, K., Niedzwiedz, C., & Platt, S. (2013). Suicide Clusters: A Review of Risk Factors and Mechanisms. *Suicide and Life-Threatening Behavior*, *43*(1), 97–108. https://doi.org/10.1111/j.1943-278X.2012.00130.x

Jashinsky, J., Burton, S. H., Hanson, C. L., West, J., Giraud-Carrier, C., Barnes, M. D., & Argyle, T. (2014). Tracking suicide risk factors through Twitter in the US. *Crisis*. Retrieved from http://econtent.hogrefe.com/doi/full/10.1027/0227-5910/a000234

Joiner, T. E. (1999). The clustering and contagion of suicide. *Current Directions in Psychological Science*, *8*(3), 89–92.

Joiner, T. E. (2003). Contagion of suicidal symptoms as a function of assortative relating and shared relationship stress in college roommates. *Journal of Adolescence*, *26*(4), 495–504. https://doi.org/10.1016/S0140-1971(02)00133-1

Kern, M. L., Park, G., Eichstaedt, J. C., Andrew, H., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining Insights From Social Media Language: Methodologies and Challenges. *Psychological Methods*, No Pagination Specified. https://doi.org/10.1037/met0000091

Knox, K. L., Conwell, Y., & Caine, E. D. (2004). If Suicide Is a Public Health Problem, What Are We Doing to Prevent It? *American Journal of Public Health*, *94*(1), 37–45.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, *70*(6), 543.

Luxton, D. D., June, J. D., & Fairall, J. M. (2012). Social Media and Suicide: A Public Health Perspective. *American Journal of Public Health*, *102*(Suppl 2), S195–S200. https://doi.org/10.2105/AJPH.2011.300608

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., & Rosenquist, J. N. (2011). Understanding the Demographics of Twitter Users. *ICWSM*, *11*, 5th.

Newman, M. (2002). Assortative mixing in networks. *Physical Review Letters*, *89*(20), 208701.

Newman, M. (2003). Mixing patterns in networks. *Physical Review E*, *67*(2), 026126.

Newman, M. (2010). *Networks: An Introduction* (1 edition). Oxford ; New York: Oxford University Press.

Newman, M., & Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E*, *68*(3), 036122. https://doi.org/10.1103/PhysRevE.68.036122

Nock, M. K., Borges, G., Bromet, E. J., Cha, C. B., Kessler, R. C., & Lee, S. (2008). Suicide and Suicidal Behavior. *Epidemiologic Reviews*, *30*(1), 133–154. https://doi.org/10.1093/epirev/mxn002

O’Dea, B., Wan, S., Batterham, P. J., Calear, A. L., Paris, C., & Christensen, H. (2015). Detecting suicidality on Twitter. *Internet Interventions*, *2*(2), 183–188. https://doi.org/10.1016/j.invent.2015.03.005

Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., … Brew, C. (2012). Sentiment Analysis of Suicide Notes: A Shared Task. *Biomedical Informatics Insights*, *5*(Suppl 1), 3. https://doi.org/10.4137/BII.S9042

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Randall, J. R., Nickel, N. C., & Colman, I. (2015). Contagion from peer suicidal behavior in a representative sample of American adolescents. *Journal of Affective Disorders*, *186*, 219–225. https://doi.org/10.1016/j.jad.2015.07.001

Ribeiro, J. D., Franklin, J. C., Fox, K. R., Bentley, K. H., Kleiman, E. M., Chang, B. P., & Nock, M. K. (2016). Self-injurious thoughts and behaviors as risk factors for future suicide ideation, attempts, and death: a meta-analysis of longitudinal studies. *Psychological Medicine*, *46*(02), 225–236. https://doi.org/10.1017/S0033291715001804

Robinson, J., Cox, G., Bailey, E., Hetrick, S., Rodrigues, M., Fisher, S., & Herrman, H. (2016). Social media and suicide prevention: a systematic review. *Early Intervention in Psychiatry*, *10*(2), 103–121. https://doi.org/10.1111/eip.12229

Russell, M. A. (2013). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More* (2 edition). O'Reilly Media.

Sanderson, D. (2009). *Programming google app engine: build and run scalable web apps on google's infrastructure*. O'Reilly Media, Inc. Retrieved from https://books.google.com/books?hl=en&lr=&id=6cL_kCZ4NJ4C&oi=fnd&pg=PR7&dq=mobile+app+scalable&ots=sKhhMVMZfh&sig=_3qSzEYQK-Q2cyD_erRlPv-jP2o

Shalizi, C. R., & Thomas, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research*, *40*(2), 211–239.

Travers, J., & Milgram, S. (1967). The small world problem. *Phychology Today*, *1*, 61–67.

Twitter. (2016a). Company | About. Retrieved May 1, 2016, from https://about.twitter.com/company

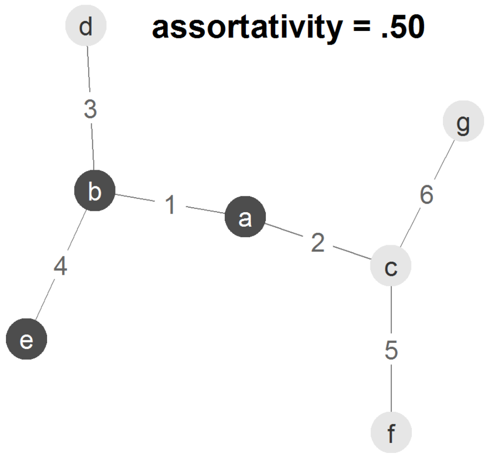
Twitter. (2016b). Getting started with Twitter. Retrieved July 5, 2016, from https://support.twitter.com/articles/215585

Twitter. (2016c). Twitter Developers. Retrieved May 1, 2016, from https://dev.twitter.com/

Wood, A., Shiffman, J., Leary, R., & Coppersmith, G. (2016). Language signals preceding suicide attempts. In *CHI 2016 Computing and Mental Health Workshop, San Jose, CA*. Retrieved from http://chi2016mentalhealth.media.mit.edu/wp-content/uploads/sites/46/2016/04/language_signals_preceding_suicide_attempts.pdf

World Health Organization. (2012). *Public health action for the prevention of suicide: a framework*. World Health Organization. Retrieved from http://www.who.int/iris/handle/10665/75166

Figures



Link Num	Node 1 ID	Node 2 ID	Node 1 Group	Node 2 Group
1	a	b	0	0
2	a	c	0	1
3	b	d	0	1
4	b	e	0	0
5	c	f	1	1
6	c	g	1	1
				r = .50

Figure 1. Illustration of assortativity as correlation of node attributes across links. The left panel depicts an example 7-node / 6-link network, in which each node belongs to one of two groups (0 = dark grey, 1 = light grey). Nodes are labelled with arbitrary letter IDs and links are labelled with arbitrary number IDs. Group membership is assortative in this network (*r* = .50), with 60% of links holding like-type nodes on either end. The right panel shows the same network, now represented as an ‘edge-list,’ in which each row is a link and each column is an attribute of that link (e.g., its link ID number) or of the nodes residing on either of its ends. Note that when a network is represented in this way, the Pearson correlation between the two columns of a given node attribute – group membership in this case – is identical to the assortativity of that attribute in the network (with some specific exceptions).

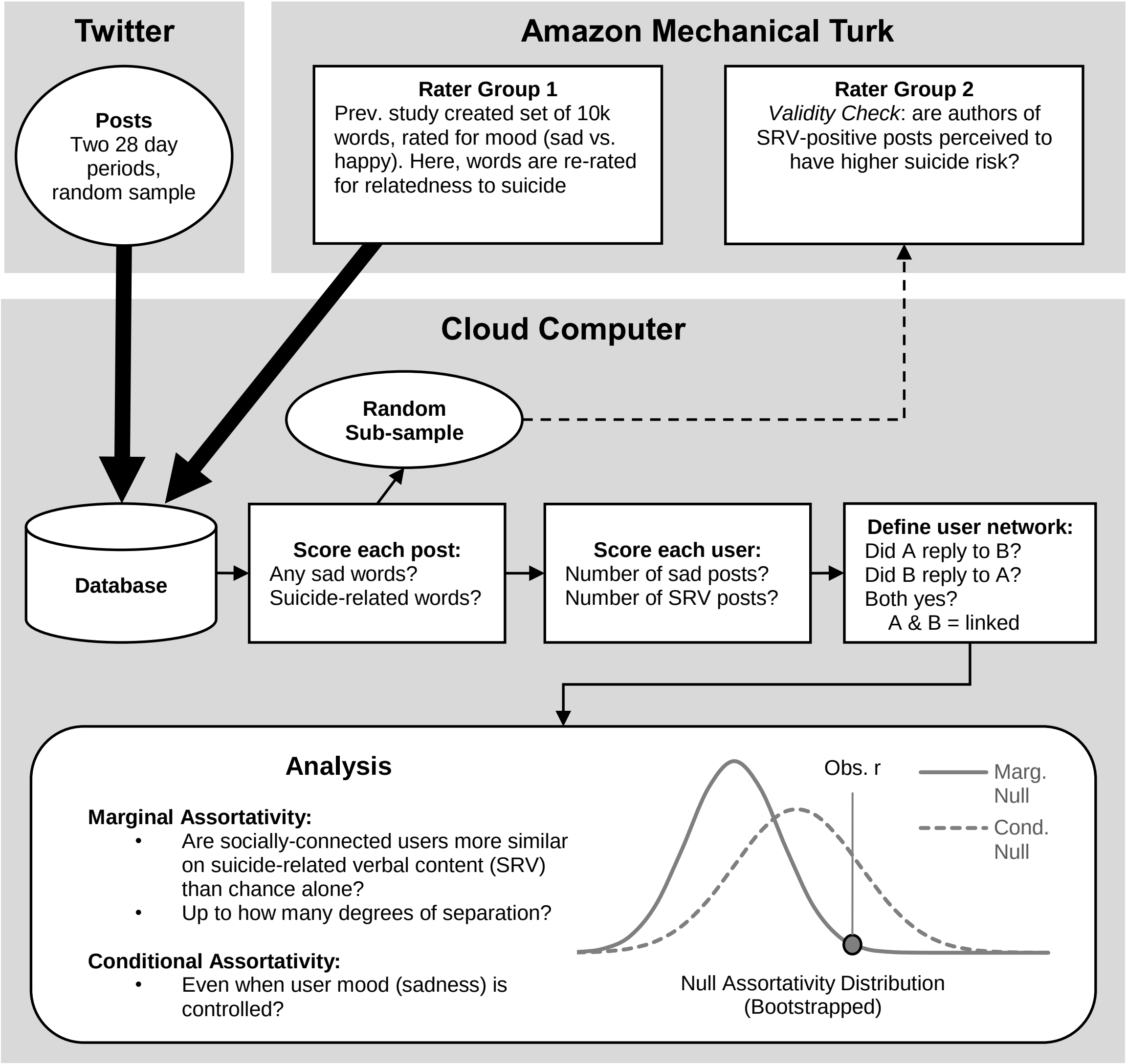

Twitter
Posts
Two 28 day periods, random sample
Amazon Mechanical Turk
Rater Group 1
Prev. study created set of 10k words, rated for mood (sad vs. happy). Here, words are re-rated for relatedness to suicide
Rater Group 2
Validity Check: are authors of SRV-positive posts perceived to have higher suicide risk?
Cloud Computer
Random Sub-sample
Database
Score each post:
Any sad words?
Suicide-related words?
Score each user:
Number of sad posts?
Number of SRV posts?
Define user network:
Did A reply to B?
Did B reply to A?
Both yes?
A & B = linked
Analysis
Marginal Assortativity:
Are socially-connected users more similar on suicide-related verbal content (SRV) than chance alone?
Up to how many degrees of separation?
Conditional Assortativity:
Even when user mood (sadness) is controlled?
Obs. r
Marg. Null
Cond. Null
Null Assortativity Distribution (Bootstrapped)

Figure 2. Diagram of study process. Investigation begins with the collection of two discontinuous 28-day periods of Twitter posts (a random sample of all posts). Simultaneously, MTurk Group 1 re-rated an existing set of 10,000 English words for relatedness to suicide (MTurk Group 1). Word ratings were used to define sad posts and posts with suicide-related verbal content (SRV). Posts were then scored by machine, but a random subsample were also re-evaluated by humans (MTurk Group 2, blind to machine rating). Users were assigned total scores base on their post scores and linked to other users with whom they mutually interacted to establish a network. Data were analyzed as described in the Methods section. Note, Obs. r = observed SRV assortativity value for the observed network.

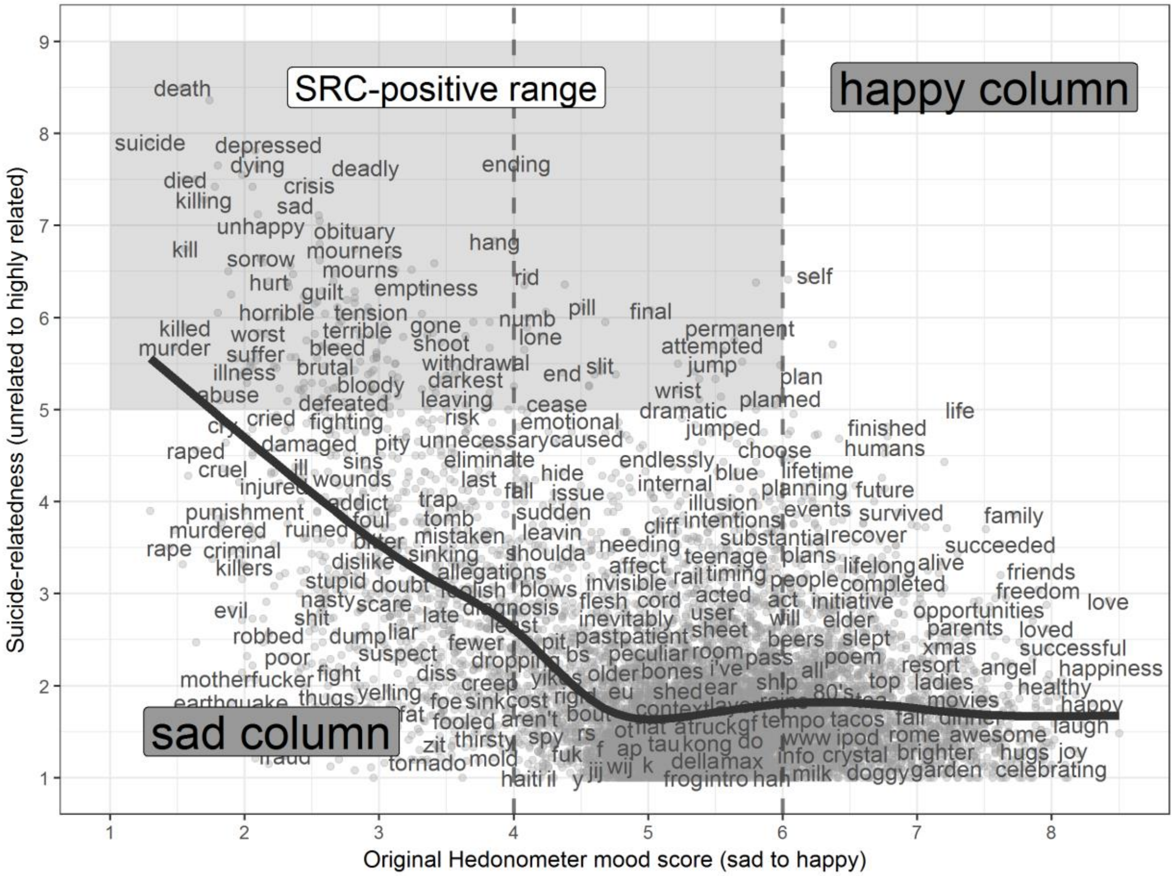


Figure 3. Mood and suicide-relatedness of the 10,222 most common words in contemporary English. Original mood scores are plotted along the horizontal axis and the suicide-relatedness ratings from the current study are plotted along the vertical axis. Each word is represented by a grey dot, a random sub-sample of which also have text displayed for the reader. The curved solid line represents the smoothed relationship between mood scores (sad to happy) and suicide-relatedness scores among words. The vertical dashed-lines demarcate sad and happy words (mood scores below 4 or above 6, respectively). The shaded box represents the set of words leading to a positive SRC score for a post. Application of these scores is further described in the Methods section and illustrated for hypothetical posts in Table 1.

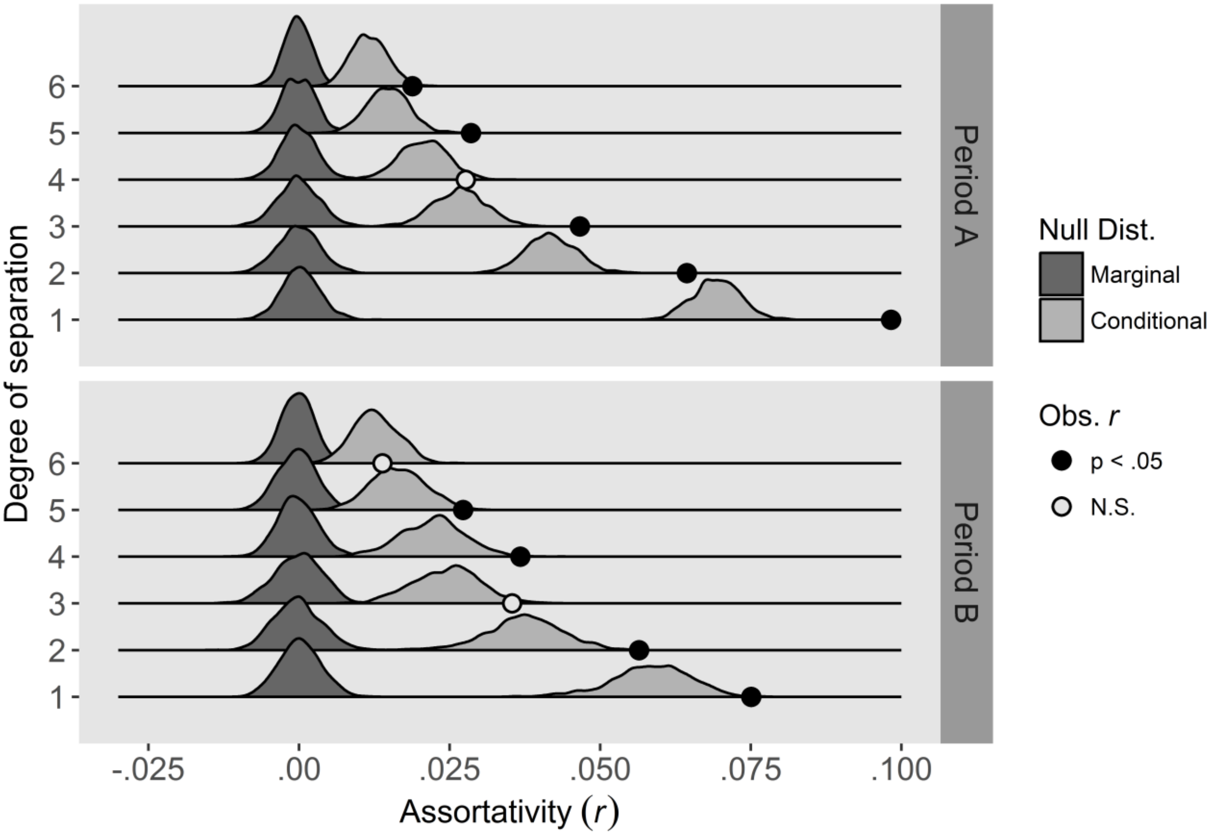


Figure 4. Observed assortativity (*r*) up to six degrees of separation and across two periods. Each point represents an assortativity estimate (horizontal axis) for the network at a given degree of separation (vertical axis). The bootstrapped null distributions for both marginal SRV assortativity (dark) and conditional SRV assortativity (light) are also provided as benchmarks for evaluating the statistical significance of those sample assortativity estimates. Consistent with traditional hypothesis testing, a sample assortativity estimate is significant when it occupies a low density (low probability) region of its null distribution. In this case, all observed assortativity estimates are significantly different than the marginal null (complete randomness), so the significance values depicted by hollow or filled points illustrate only whether the observed assortativity was significantly different from the conditional null (randomness conditional on sadness), at a given degree of separation. Note, because marginal null distributions are completely random, they will always center around 0 assortativity; however, because the conditional null distributions account for the assortativity in sadness (which is known to decline as a function of social distance), these distributions will center at difference values. Specifically, they will be centered on the most plausible SRC assortativity value one would expect, knowing the sadness scores of everyone in the network.

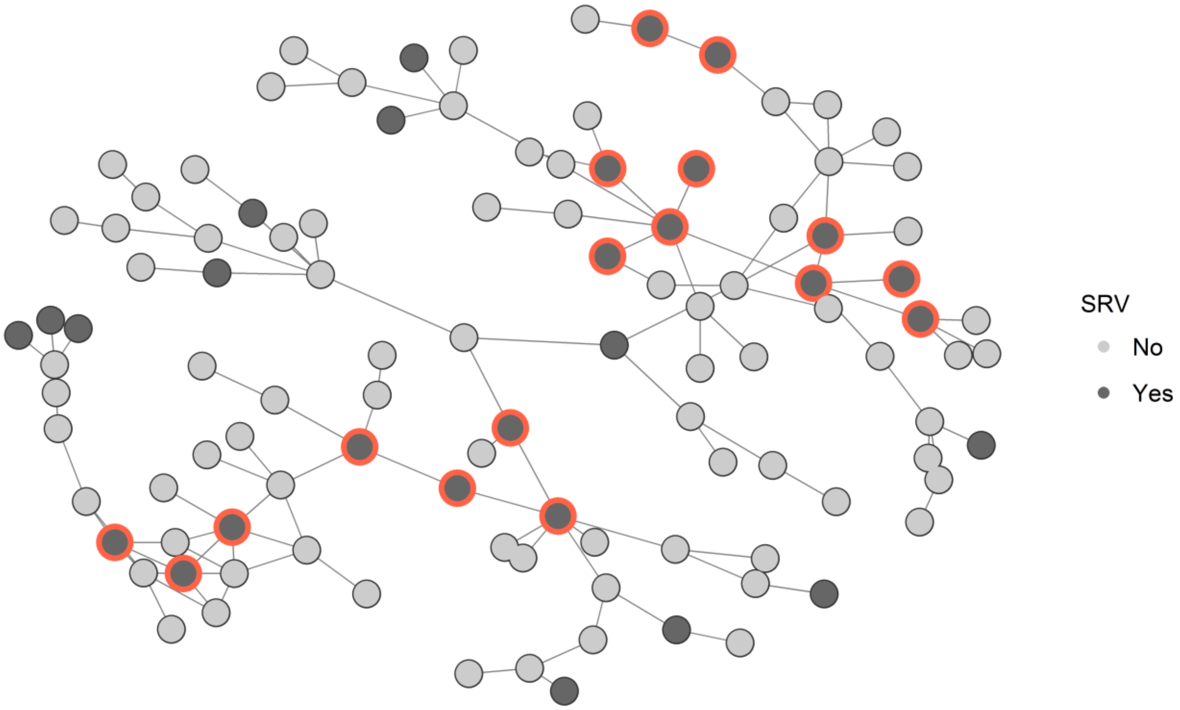


Figure 5. Visual demonstration of actual SRV assortativity in a subsample of nodes from the first study period. Light circles represent nodes with no SRV-positive posts during the first study period; dark circles represent nodes with one or more SRV-positive posts during that period. Thick red outlines highlight the SRV-positive nodes that are linked to at least one other SRV-positive node in the subsample. In this subsample, 59% of SRV-positive nodes are directly connected to another SRV-positive node; likewise, 54% of SRV-negative nodes are directly connected to at least one SRV-positive node.

Tables

Table 1. Example posts with application of scoring rules from Methods

		Sad	Happy	Suicide-related			User	User SRV
User	Text	Word	Word	Word	Sad Post	SRV Post	Sadness	Score
A	"I'm so sad! Gonna kill myself"	sad, kill		kill	1	1	**3**	**2**
A	"I'm the worst lol :)"	worst	lol, :)	worst	1	0		
A	"My final day on Earth…"			final	0	1		
A	"Just got in a fight"	fight			1	0		
B	"It's a sad day"	sad			1	0	**1**	**0**
B	"I love my life"		love, life		0	0		

Note. User represents the ID of a hypothetical Twitter user, whose posts are given in the Text column. The Sad Word, Happy Word, and Suicide-related Word columns identify which words (if any) from the user's post fall into the sad, happy, or suicide-related word categories described in the Methods section and depicted in Figure XXX. Sad Post and SRV Post represent dummy coded variables representing whether the post would be classified as sad (scored 1 if the post included any sad word; 0 otherwise) or SRV-positive (scored 1 if the post included a suicide-related word, but no happy words; 0 otherwise). User Sadness and User SRV Score columns represent the total number of sad and SRV-positive posts by each hypothetical user; these are the scores being analyzed in this investigation (i.e., Are SRV scores more assortative than chance? Do they remain so after controlling for sadness scores?).