Can x2vec save lives? Integrating graph and language embeddings for automatic mental health classification

Alexander Ruch
Cornell University, Ithaca, NY 14850, United States of America
Graphika New York, NY, United States of America
E-mail: amr442@cornell.edu and alex.ruch@graphika.com

Keywords: graph embeddings, language embeddings, information networks, social media, mental health

Abstract
Graph and language embedding models are becoming commonplace in large scale analyses given their ability to represent complex sparse data densely in low-dimensional space. Integrating these models’ complementary relational and communicative data may be especially helpful if predicting rare events or classifying members of hidden populations—tasks requiring huge and sparse datasets for generalizable analyses. For example, due to social stigma and comorbidities, mental health support groups often form in amorphous online groups. Predicting suicidality among individuals in these settings using standard network analyses is prohibitive due to resource limits (e.g., memory), and adding auxiliary data like text to such models exacerbates complexity- and sparsity-related issues. Here, I show how merging graph and language embedding models (metapath2vec and doc2vec) avoids these limits and extracts unsupervised clustering data without domain expertise or feature engineering. Graph and language distances to a suicide support group have little correlation ($\rho < 0.23$), implying the two models are not embedding redundant information. When used separately to predict suicidality among individuals, graph and language data generate relatively accurate results (69% and 76%, respectively) but have moderately large false-positive (25% and 21%, respectively) and false-negative (38% and 27%, respectively) rates; however, when integrated, both data produce highly accurate predictions (90%, with 10% false-positives and 12% false-negatives). Visualizing graph embeddings annotated with predictions of potentially suicidal individuals shows the integrated model could classify such individuals even if they are positioned far from the support group. These results extend research on the importance of simultaneously analyzing behavior and language in massive networks and efforts to integrate embedding models for different kinds of data when predicting and classifying, particularly when they involve rare events.

1. Introduction

Complex social dynamics like behavior and communication patterns are often sparse in the sense that such data for any given individual, group, or entity have few observations of interest and many null or irrelevant observations. For example, in a network of 1000000 individuals, someone who has interacted with just 1% of the network has large number of ties to others (10 000), but their vector of actual and potential ties to others is nearly empty. The same outcome develops from a text that includes 100 unique words from a vocabulary of 10 000 words. This sparsity compounds the already complex practices of network analysis and natural language processing. Embedding models, however, are useful solutions to this problem, as they effectively (in terms of time and memory) learn dense low-dimensional representations of complex sparse data. In other words, embedding models can represent the aforementioned individual’s network connectivity in a 128-dimension vector with non-null elements instead of a mostly empty 1000 000-dimension vector.

Embedding models were first developed in natural language processing to represent similarities between words given the similar linguistic contexts in which they co-appear (e.g., word2vec in Mikolov et al 2013).
These models were then extended to embed words using their subword information (positioning 'write' close to 'writes' and 'wrote') and extended to embed sentences and documents based on the words and topics in them (Bojanowski et al 2017, Le and Mikolov 2014). Models to embed networks were based on these techniques, treating the nodes in a graph like words in a document. Nodes are sampled from the graph using random walks (Perozzi et al 2014), and the random walk algorithm can be adjusted to account for homophily and structural equivalence (Grover and Leskovec 2016) and for heterogeneous/multi-modal graph structure (Dong et al 2017). Collectively, these vectorized embedding models are often referred to as 'x2vec' models (e.g., Grohe 2020), as their names often end in '2vec' (e.g., word2vec, doc2vec, node2vec, metapath2vec).

Given their facility for analyzing complex network and language problems, integrating these two complementary embeddings may overcome the limitations of any single embedding and extend computational social science by making possible simultaneous analysis of multiple types of action. For example, integrated models may be especially helpful if one’s research involves predicting rare events or classifying members of hidden populations, tasks requiring huge and sparse datasets for generalizable models. Using standard network analyses in these cases can be prohibitive due to resource limits (e.g., memory), and adding auxiliary data like text to complement these models further exacerbates their complexity and sparsity-related issues, as one sparse matrix combined with another often produces an even larger sparse matrix. For example, suicidality is stigmatized worldwide and frequently develops along with a wide range of other physical and mental health comorbidities (Corrigan 2004, Corrigan and Watson 2002). Individuals struggling with suicidality often seek sporadic support in amorphous online groups, making it difficult to classify members of these groups and predict those who may become suicidal. Producing representative models of these outcomes requires lots of sparse data, in other words, as most individuals never post in a support group and those who do may post in a large number of other online groups before they finally post in a support group.

In this paper, I show how integrating graph and language embedding models (metapath2vec and doc2vec—two of the aforementioned ‘x2vec’ models) avoids the aforementioned limits while extracting unsupervised clustering data on networks and language without requiring domain expertise. I then measure how correlated these embeddings’ distances are with each other to evaluate the extent to which they are extracting redundant similarity data. Next, I compare graph and language embeddings’ separated and integrated abilities to generate accurate predictions of rare events: posting submissions to Reddit’s r/SuicideWatch subreddit for suicide support in a heterogeneous network of 35.5 million nodes and 190 million edges. Finally, I quantitatively and qualitatively evaluate the integrated model’s ability to avoid false-positives and false-negatives in this context—which is important when there are vastly more negative cases than positive cases, as a model could have apparently high accuracy by always predicting ‘negative’. The results of this study extend research on the importance of simultaneously analyzing behavior and language in massive networks and efforts to integrate different kinds of embedding models for prediction and classification tasks, especially when they involve rare events. Note that while I use two specific ‘x2vec’-style embedding models for my application, researchers and data scientists working on similar tasks involving predicting/classifying rare events with complex, unlabeled data may use other unsupervised embedding models with success. The key point to highlight here is the utility of using these kinds of models for complex tasks involving both network and language data.

All data used in this study are publicly available from Reddit. The code used to sample the network and perform graph and language embeddings, natural language processing, and machine learning predictions is available at https://github.com/AlexMRuch/Can-x2vec-Save-Lives. Care is taken to avoid presenting personally identifiable information in this paper, and all examples of submissions are paraphrased. This work was approved by the appropriate institutional review board of the author. Please note this work does not make diagnostic claims related to mental illness or suicide.

2. Related work

2.1. Network analysis

Similarity between nodes in a network can be measured in various ways; however, many methods focus on extracting nodes’ structural equivalence: the extent to which two or more nodes share the same set of edges to other nodes in the network (Borgatti and Grosser 2015). These nodes hold the same position in a network and are equivalent to each other even if they share no edge(s) between themselves. For example, if a company had two managers who worked with the same employees but who did not work with each other, these managers would have structurally equivalent positions in the company’s network. Moreover, since these managers are connected to all of the same people, they can be said to be redundant in the network and may replace one another, as neither works or interacts with anyone else in the network who the other does not also work or interact. In many cases, however, we do not wish to know if two nodes share exactly the same set of neighbors.
Methods of measuring nodes' similarities can either derive continuous distances between nodes or identify classes of nodes that are more or less similar (e.g., for clustering or community detection). For example, continuous measures include the following methods. Euclidean distance equals the number of network neighbors that two nodes do not share and is normalized by dividing by the total number of nodes in the network. The Pearson correlation coefficient can also be used to get a normalized measure of shared network neighbors. The Jaccard coefficient equals the intersection of network neighbors shared divided by the expected number of common neighbors in a random network—is not biased in large networks divided by the geometric mean of those nodes' degrees. Shi and Macy (2016) find that each of these methods are biased when measuring similarity among nodes in large networks, as the nodes in these networks can have highly skewed degree distributions which can overemphasize the effect of nodes’ out-degrees on their similarity in many of these methods. They recommend using the standardized coincidence ratio in such cases, as its measure of similarity—based on the number of observed common neighbors divided by the expected number of common neighbors in a random network—is not biased in large networks with millions of nodes.

By contrast, similarity measures that identify discrete groups of similar nodes classify sets of nodes that share many connections within their respective groups but have little or no connection to other groups. Such groups are often called cohesive subgroups (Frank (1995), (1996)). For example, using hierarchical agglomerative clustering on a network’s similarity matrix (generated from any of the metrics in the previous paragraph) produces a tree-like structure of groups of increasingly similar nodes from pairs of nodes to small groups of nodes to larger groups of nodes all the way to where the whole network is clustered into one group. One can use a dendrogram to visualize this tree and select a cutoff threshold for similarity to generate classes of nodes that formed within the threshold. Modularity is a similar algorithm that maximizes the proportion of edges lying between nodes in a group minus the expected proportion of edges existing between nodes in a group at random. Unfortunately, this method can suffer from a ‘resolution limit’ in which relatively small classes of similar nodes are overlooked by the algorithm (Fortunato and Barth´e 2007).

Hierarchical stochastic block models (HSBMs) attempt to bridge the gap between continuous and discrete measures of similarity. In HSBMs, levels of increasingly similar groups (or ‘blocks’) of nodes are produced using algorithms that iteratively partition nodes into groups in which member nodes share many connections to each other while sharing no or few connections to nodes in other groups. These models do not suffer from the ‘resolution limit’ that hampers modularity measures (Peixoto 2014). Additionally, these models can adjust nodes’ similarities based on their degree and can generate overlapping (i.e., continuous) groups as well. Lastly, HSBMs can be used to analyze core-periphery structures in networks and to predict whether edges exist between nodes. Despite these advances, HSBMs are computationally expensive and may be unsuccessful in large networks, as their algorithms require keeping a large amount of data in memory (Fortunato and Hric 2016).

Given the complexity of accurately measuring network similarity at scale, attempts to add auxiliary language data to network analyses have been limited—yet, such data is recognized as important to accurately capturing complex social interactions in ways that extend beyond what either behavior or communication data captures on its own (Benamara et al 2018). Many studies that have integrated these two forms of data have used qualitative discourse analysis of text, which—while powerful at capturing nuanced meanings in interactions—are limited to small scale studies and require expert working knowledge of the given context (Moser et al 2013). On the other hand, many studies that have used quantitative text analysis approaches, like natural language processing, in addition to network analysis have often used few measures of network structure and similarity (Benamara et al 2018, Gonzalez-Bailon et al 2010). Bail’s work (2016) contrasts with these trends by directly integrating language similarity data with network structure measures (e.g., closeness and betweenness centrality) in his study of autism spectrum disorder advocacy organizations and their communication practices. Though his method enabled the analysis of cultural bridges through a combined behavior and communication framework, his approach to natural language processing used a bag-of-words tokenization that removed much of the language’s context that comes from how words are used in a particular order (e.g., ‘I know, vaccines do not cause autism’ vs ‘I do not know, vaccines cause autism’). Leaving such properties of language out of our analyses may cause us to miss important dynamics in modeling social interaction and similarity.

2.2. Embedding models
2.2.1. Graph embedding
The similarity measures mentioned above run on networks’ adjacency matrices. If a network has 1,000,000 nodes, the size of this matrix will be 1,000,000^2 (i.e., it will contain one trillion elements). If the network is sparse, the elements in this adjacency matrix will be mostly empty, contributing to substantial computational overhead that generates little information. Embedding models avoid these limitations by learning dense
low-dimensional representations of data. In other words, one can use embedding models to learn similarities between nodes in this network of 1,000,000 nodes without analyzing its adjacency matrix and produce a matrix of 1,000,000 rows (one for each node) and \( D \) columns, where \( D \) represents the depth of similarity dimensions representing the network (typically 64, 128, 256). The positional values of these \( D \) elements will be rarely be empty, which addresses our sparsity concern, and the fact that \( D \) is miniscule in size compared to the number of nodes in the network allows us to avoid memory limits. This embedding matrix can then be used for other downstream similarity tasks like cluster analysis.

Graph embeddings can be constructed through three methods, each of which has its own strengths and limits (Cai et al. 2017, Goyal and Ferrara 2017). First, factorization of graph adjacency matrices (Ahmed et al. 2013) can extract similarities between nodes and their neighbors (first order proximity) and similarities between nodes and their neighbors’ neighbors (second order proximity), as is done in LINE (Tang et al. 2015). This approach is limited, however, as it must run on networks’ adjacency matrices, which hampers its ability to run on networks that are too large to load into memory. HOPE (Ou et al. 2016) attempts to overcome this limit by using singular value decomposition to create a low-dimension representation of the adjacency matrix and then embeds first and second order proximities in the similarity matrix.

While HOPE does reduce memory load and embedding’s computational complexity, factorization-based embedding methods still require all nodes in the network to be loaded into memory at once. Many deep learning embedding methods have the same requirement. For example, SDNE (Wang et al. 2013) passes nodes and their neighbors through a series of autoencoders to learn the graph’s structure. This method is both computationally efficient and capable of learning non-linear features of structural similarity without supervision; however, because the whole network must be loaded in memory at once, it cannot be used in cases where the target network cannot fit in memory or where it is not possible to collect the whole network (e.g., when nodes and edges may be missing or cannot be extracted due to sampling limitations or restrictions).

Using random walk sampling methods in deep learning embedding models avoids the space limit issue of models that require loading the whole network into memory. Many of these models learn nodes’ primary and secondary similarities by iterating through sequences of nodes extracted from random samples and using a skip-gram-based model to generate embeddings that maximize the probability of observing the neighbors of nodes—up to a specific window length—conditional on nodes’ present embeddings. For example, if a walk from a sample included the nodes \( \text{user}_1, \text{user}_2, \text{user}_3, \text{user}_4, \text{user}_5 \), and so on, and if nodes are embedded over a window of 2, then the skip-gram model would increase \( \text{user}_1 \)’s similarity to \( \text{user}_2 \) and \( \text{user}_3 \), then increase \( \text{user}_2 \)’s similarity to \( \text{user}_1, \text{user}_3, \text{and} \text{user}_4 \), then increase \( \text{user}_3 \)’s similarity to \( \text{user}_1, \text{user}_2, \text{user}_4, \text{and} \text{user}_5 \)—and so on. To quicken training, the skip-gram model is often run jointly with a negative sampling process that adds distance between the focal node and a set of non-neighbor nodes. The details of this algorithm and how it is applied in the present paper are outlined in section 3.2.

DeepWalk (Perozzi et al. 2014) was the first graph embedding model to use the skip-gram model. It used strictly random walks to generate samples of nodes to train the embedding model. Grover and Leskovec (2016) also use skip-gram to embed nodes, but they use two parameters to bias the random walk’s probability of returning to the node last sampled and to bias the walk’s probability of sampling nodes close to the node last sampled versus that are nodes far from it. Together with the skip-gram model, this biased random walk sampling process helps embed nodes that can have more or less homophily and structural equivalence. The metapath2vec model (Dong et al. 2017) extends this approach for embedding heterogeneous graphs comprised of different types of nodes (e.g., universities, researchers, papers, and conferences). To do this, one defines a metapath-biased random walk scheme: a set of node types to sample in a specific order (e.g., \{researcher, paper, conference, paper, researcher\} embeds the similarities of researchers based on the conferences in which they both present papers). A variant, metapath2vec++, modifies the negative sampling procedure as well to generate a set of embeddings that are specific to each type of node encountered (e.g., researchers are embedded near researchers and far from papers, which are embedded closely).

2.2.2. Language embedding

The skip-gram model and negative sampling process were originally developed to embed words using word2vec (Mikolov et al. 2013). In the context of language, this embedding model learns to maximize the probability of words that co-occur. For example, if you read the sentence ‘I drink _____ for breakfast’, words like coffee, tea, milk, juice, and water probably come to mind based on how you have likely heard or read this same sentence with each of these words in the middle many times before. In other words, each of these breakfast beverages shares a similar language context. One limitation of this model, however, is that it does not embed information about word order. For example, it cannot estimate the meaning and similarity of
Moreover, it cannot deduce the fact that people from Buffalo, NY, only call them ‘wings’. Homonyms like ‘Buffalo buffalo Buffalo buffalo’ to other words by simply combining the embeddings of ‘buffalo’ and ‘wings’1. The word2vec sampling algorithm can be updated to identify and embed such phrases, however (Mikolov et al 2013). This approach has since been modified to embed subword information (Bojanowski et al 2017) as well as complete sentences and documents (Le and Mikolov 2014). The latter method of embedding documents with doc2vec (Le and Mikolov 2014) is particularly useful, as it is able to learn to represent semantic as well as syntactic information about language. The details of the doc2vec algorithm and how its two variations are applied in the present paper are outlined in section 3.2.

2.2.3. Multitask embeddings

Algorithms that integrate different embedding approaches are called multitask embedding models. By combining representations of different data forms, these models greatly enhance embedding’s performance on prediction and classification tasks. For example, Yang et al (2015) use DeepWalk and matrix factorization respectively to embed homogeneous graphs with auxiliary text data. Their text-associated DeepWalk model achieved state-of-the-art node classification performance and worked very well in sparse and noisy networks with small training samples. Zhang et al (2017) build on this approach by adding homophily, topology structure, and text data to their embedding model’s learning objective function. These data together generate embeddings that better represent the complex interrelations between nodes’ local and global structure and the different language contexts over which they exist. While this model also performs well on multi-node classification, it is limited to homogeneous networks and uses a bag-of-words approach to embedding language that does not embed the syntactic and structural information in documents. An et al (2018) use a similar method to embed networks with auxiliary data that uses triples to distinguish between the different contexts in which two nodes may be related, therefore allowing it to embed heterogeneous knowledge graphs. This model greatly improves the embeddings’ quality; however, it too stops at embedding words in nodes’ textual context instead of learning to represent the entire text. Xiao et al (2017) develop a partial solution to this problem of embedding texts in multitask embeddings by averaging texts’ word embedding vectors. While this step improves the representations of more complex semantics in texts, it does not embed the larger structural features of documents that may be important to differentiating the language context of nodes.

2.3. Suicidality

People who struggle with mental health are often labeled as abnormal and are stigmatized. These labels and stresses of stigmatization, including stereotypes and discrimination, can exacerbate the negative effects of one’s mental health status (Link et al 1989). To cope, many who struggle with mental health avoid discussing it or seeking professional help or social support (Corrigan 2004). Individuals may also withdraw from social life and interaction as they self-stigmatize and try to avoid rejection (Corrigan and Watson 2002). Efforts to reduce mental health stigma find negative attitudes and prejudice are pervasive regardless of age, education, and whether people know others who struggle with mental health (Crisp et al 2000). Even after such interventions, although levels of stigma decrease among some conditions like depression and anxiety, stigma is still high among the population—especially among teens and young adults (Crisp et al 2005).

Like mental health illness, suicidality—serious thoughts, plans, and attempts related to suicide—is particularly stigmatized, and epidemiological research finds that 90% of people who attempted suicide had a diagnosable mental health disorder prior to their attempt (Schreiber and Culpepper 2019). This is concerning, as 47 000 individuals in the US commit suicide yearly, making it the 7th leading cause of years of life lost (Schreiber and Culpepper 2019). Moreover, these risks are 5–10 times higher for youth and adolescents (Kennebeck and Bonin 2019).

Computational social scientists have studied these dynamics in a variety of contexts. For example, natural language processing has been used to classify depression, post-traumatic stress disorder, attention-deficit/hyperactivity disorder, and other mental health qualities in both clinical and non-clinical settings (Cook et al 2016, Gkotsis et al 2017, Gundlapalli et al 2013, Seitz and Kristy 2016). Similarly, De Choudhury et al (2017) find mental health support groups are highly active online and that esteem-boosting comments and network support from other group members lowered one’s risk of posting increasingly suicidal submissions in the future. Unfortunately, other work in this setting also finds that significant inequalities persist in regard to which types of posts receive the greatest attention, with dissociative posts receiving the fewest comments and psychotic posts getting the most (Ruch et al 2019). Posts presenting with depression and anxiety were neither more nor less likely to receive supportive comments from the community compared to posts labeled with any other mental health condition.

1 Moreover, it cannot deduce the fact that people from Buffalo, NY, only call them ‘wings’. Homonyms like ‘Buffalo buffalo Buffalo buffalo buffalo Buffalo buffalo’ are even more complex and out of scope: https://en.wikipedia.org/wiki/Buffalo_buffalo_Buffalo_buffalo_buffalo_buffalo_buffalo_Buffalo_buffalo.
3. Data and methods

3.1. Data and preprocessing

To examine whether graph and language embeddings can be used to predict potentially suicidal individuals in online settings, I used submission data from Reddit—one of the largest social news aggregation, web content rating, and discussion websites. I extracted all public data available on Reddit from June 2005 to June 2018 from pushshift.io, a Reddit-based data project that stores and analyzes Reddit data in real time and releases monthly data dumps of all the information gathered. From this data, I constructed a database of 490 million submissions with 4.3 trillion comments from 66 million authors in 27 million subreddits. Data has a heterogeneous (multimodal) network structure, with authors having links to submissions and comments, comments having links to submissions, and submissions having links to subreddits. All data includes timestamps of when the information was first posted to Reddit, and all data in the database was collected at the moment it was first posted on Reddit—and so all data exists in its original unedited form. The structure of this heterogeneous network is presented in figure 1. Please note that table 1 lists and defines a set of acronyms that appear frequently throughout this paper.

I used a series of forest fire network sampling models to collect one main SuicideWatch (SW) sample and three complementary samples with author seeds from mental health (MH), self-help (SH), and random (R) subreddits.

Forest fire models generate network samples by either starting from a set of seed nodes or random nodes and then ‘burning’ (sampling) every neighboring node of the node set last sampled with a predetermined ‘burn’ probability. If this probability is set to 1, then the forest fire model will sample the same as breadth first search. If the ‘fire’ dies out before the predetermined number of nodes is sampled, then the ‘fire’ can restart from an already sampled node and attempt to ‘reburn’ previously unsampled nodes. For sampling large graphs, Leskovec and Faloutsos (2006) demonstrate that the forest fire sampling method is one of the most accurate techniques for recovering the ‘back-in-time’ structure of temporal and evolving networks. Kurant et al (2010), however, note that this method—and every other method based on random walks—is biased toward sampling nodes with large degrees and thus generates samples that overestimate networks’ degree. The magnitude of overestimation bias decreases as the fraction of nodes in the network or network subgraph increases. Frank et al (2012), however, find that this bias is relatively low once at least 15% of the target network or subgraph is sampled.

The forest fire model I constructed sampled nodes as follows. For the main SW sample, I first identified all unique authors who posted in SW at least once and who had at least 20 submissions in total—regardless of the subreddit in which they posted it. Next, from these 24,281 unique SW submission author seeds, I ‘burnt’ 20% of nodes ($V = 4948$; $V$ stands for nodes). I then identified all of these authors’ submissions across any subreddit ($V = 777,243$; 94 submissions/author) and “burnt” 20% of them ($V = 155,646$; 31 submissions/author). I identified all of the comments these authors made ($V = 9,611,359$; 1942 comments/author) and “burnt” 20% of them ($V = 1,415,357$; 286 comments/author)—making sure to burn no more than 1 comment per unique submission to avoid sampling multiple comments authors may post on the same submission. In this process, I also ‘burnt’ the submission and submission author for which the ‘burnt’ comment was posted. I next identified all comments that were posted to the ‘burnt’ submissions ($V = 447,579,856$) and ‘burnt’ 1% of non-SW authors comments ($V = 4,475,200$) and ‘burnt’ all SW authors’ comments ($V = 2,109,393$). This generated a SW sample of 6,584,593 observations ($\sim V = 4,475,200 + 2,109,393$). These observations included data for the following: ~21 K subreddits on which ~700 K submission authors (~1% of all unique Reddit authors) posted ~1.3 M submissions on which ~1.6 M comment authors posted ~6.5 M comments. Giving the sampling process of the forest fire model, the graph produced from these observations generates a connected component.

I then used identically structured forest fire models to sample 7.0 M observations from a seed list of mental health subreddits, to sample 7.6 M observations from a seed list of self-help subreddits, and to sample 5.8 M observations from 5000 randomly selected authors across all subreddits (= the random sample). In this last case, 5000 unique authors were randomly selected as seeds since this number was quite close to the number of unique authors ‘burnt’ by the first ‘fire’ in each of the previous forest fire models (e.g., 4948 unique authors were ‘burnt’ in the SW sample). These four contexts were chosen to generate representative samples of severe as well as general mental health related activity in addition to activity related to self-improvement and activity ‘normal’ to Reddit more broadly. Altogether, the four samples generated one large heterogeneous network of 35.5 M nodes ($V$) and 190 M edges ($E$), which was also a connected component as few of the ‘burnt’ authors

---

2 Mental health subreddits include mentalhealth, mentalillness, addiction, alcoholism, Anger, Anxiety, BipolarReddit, depression, domesticviolence, EatingDisorders, OCD, Phobia, psychotropicredit, ptsd, schizophrenia, survivorsofabus, rape, OpiatesRecovery, ADHD, and igtetsbetter. Self-help subreddits include selfimprovement, zenhabits, productivity, personalfinance, frugal, decidingtobebetter, GetMotivated, getdisciplined, LifeProTips, and LifeImprovement.
### Table 1. Commonly used acronyms and their definitions.

| Acronym | Definition |
|---------|------------|
| SW      | SuicideWatch, a peer support group for individuals struggling with suicidal thoughts. Please visit the subreddit’s webpage here for more details: https://www.reddit.com/r/SuicideWatch/. |
| MH      | A set of 20 mental health subreddits from across Reddit (see footnote 2). |
| SH      | A set of 10 self-help subreddits from across Reddit (see footnote 2). |
| R       | A large set of random subreddits from across Reddit. |
| MP2V    | Metapath2vec, a vectorized heterogeneous graph embedding model that embeds nodes in space independent of node type—placing nodes of the any type (e.g., authors, subreddits, submissions) close to each other based on how often those nodes appear near each other in sequences of biased random graph walks. This model was created by Dong et al (2017). |
| MP2V++  | Metapath2vec++, a vectorized heterogeneous graph embedding model that embeds nodes in space conditional on node type—placing nodes of the same type (e.g., authors) closer to each other and farther from nodes of other types (e.g., subreddits, submissions) depending on how those same-typed nodes appear near each other in sequences of biased random graph walks. This model was developed by Dong et al (2017). |
| D2V     | Doc2vec, a vectorized document embedding model that learns to encode representations of texts (submissions) as well as user-supplied submission-level metadata for texts (e.g., texts’ authors and subreddits in which texts are published). This model was made by Le and Mikolov (2014). |
| DBOW    | The distributed bag-of-words model of Doc2vec. In brief, this model learns to represent text documents (e.g., submissions) based on the topics and themes of words that are present in them. |
| DMM     | The distributed memory model of Doc2vec. In brief, this model learns to represent texts (e.g., submissions) based on the topics and themes of words that are present in them. |
| UMAP    | The uniform manifold approximation and projection model of dimensionality reduction created by McInnes et al (2018). UMAP better captures the complex non-linear relationships over local and global relationships in high-dimensional space versus linear models like principal component analysis (PCA) and T-distributed stochastic neighbor embedding (t-SNE), another popular non-linear dimension reduction model. |

had not made at least one submission or comment to one of Reddit’s popular subreddits like r/announcements, r/funny, r/AskReddit, and r/gaming—the four most followed subreddits.

#### 3.2. Graph embedding with metapath2vec

To embed this network while preserving the structural equivalence between nodes, I constructed a metapath-biased random walk sampling algorithm. The walker recursively sampled nodes over metapath context of [subreddit, submission, author, submission, subreddit] with the objective of drawing samples metapath2vec (MP2V) could analyze to extract similarities between subreddits through the authors who post submissions in them. Each submission node was walked over 1000 times for a distance of 100 steps, as recommended by Dong et al (2017).

Sampling graphs with random walks is computationally efficient in terms of time and memory. First, storing nodes’ immediate neighbors is $O(|E|)$. Retrieving nodes’ neighbors is then $O(|V|)$. Storing
interconnections between nodes’ neighbors is $O\left(a^2|V|\right)$, where $a$ is the graph’s average degree and is usually small for real-world networks (e.g., in the main sample for this study, the graph’s average degree is 9.1). Preprocessing transition probabilities makes walking from nodes $O\left(1\right)$. Writing walks’ real-time sampling results to disc instead of RAM saves a huge amount of memory. For example, the random walk sample file generated for MP2V was 50 GB; however, an HSBCM could not be run on the same network data as it quickly overflowed the server’s 128 GB of RAM. Since walks are independent, one can also parallelize the sampler using multiprocessing to greatly enhance sampling speed. This leads MP2V—including sampling and embedding models—to be $\sim 7.6$ times more efficient than hierarchical stochastic block modeling while using much less memory (see supplemental note 1 (https://stacks.iop.org/JPCOMPLEX/01/035005/mmedia)).

The heterogeneous skip-gram model MP2V used to learn dense low-dimensional representations of heterogeneous graphs functions as follows. First, heterogeneous graphs are composed of nodes of different types and are described as $G = (V, E, T)$, where $V$ is nodes, $E$ is edges, and $T$ is node type (e.g., author, comment, submission, subreddit). To embed $G$, the heterogeneous skip-gram model maximizes the probability of node $v$ having heterogeneous content $N_t(v), t \in T_V$:

$$\arg \max \theta \sum_{v \in V} \sum_{y \in \mathcal{M}(v)} \log p(c_t|v; \theta)$$

where $N_t(v)$ is $v$’s neighborhood among the $t$th type of nodes, $c_t$ is $v$’s neighboring node of type $t$, and $p(c_t|v; \theta)$ is a softmax function—otherwise known as a multinomial logistic function. The softmax function can be represented as $p(c_t|v; \theta) = \frac{e^{X_{ct}}}{\sum_{t \in T_V} e^{X_{ct}}}$, where $X_v$ is the $v$th row of $X$, the embedding vector for node $v$. If the embedding depth for MP2V is 128 dimensions, for instance, then $X_v$ will have 128 elements.

Finally, given the substantial computational complexity involved in calculating the denominator of this softmax function, Dong et al (2017) modify Mikolov et al (2013) negative sampling procedure to the heterogeneous context, drawing a small sample of size $M$ of neighbor and non-neighbor nodes to reduce the number of nodes involved in the denominator. Whereas the MP2V algorithm samples these nodes from the network independent of $v$’s node type, a variant of the metapath2vec algorithm called metapath2vec++ only samples nodes from the same node type as $v$, generating a graph representation in which nodes of the same type are embedded closely to each other and far from nodes of other types; however, because metapath2vec has been shown to outperform metapath2vec++ on many prediction and classification tasks and because the goal of this paper is to predict whether or not an author will post a submission in Reddit’s SuicideWatch subreddit, I will only use metapath2vec here (see supplemental figure 2 for a comparison).

Running MP2V on the full sample of 35.5 M nodes and 190 M edges over the subreddit, submission, author, subreddit metapath scheme generated embeddings for 1.8 M unique subreddit and author nodes. Embeddings had 128 dimensions and were based on a window size of 7 with a negative sampling size of 5. This size is fewer than 35.5 M nodes as nodes belonging to submission and comment types were not embedded, because of MP2V’s minimum appearance threshold, and since of sampling variation that may have bypassed nodes that otherwise would have surpassed MP2V’s minimum appearance thresholds. Total sampling and computation time of these processes was less than one day, and the MP2V model file generated for embeddings was only 0.9 GB. All computing was done on a 32-core 4.1 GHz 1950X Threadripper™ processor with 128 GB RAM. Code to sample metapaths and run MP2V embedding was obtained from the developer’s website: https://ericdongyx.github.io/metapath2vec/m2v.html.

### 3.3. Language embedding with doc2vec

Le and Mikolov (2014) developed what’s become known as doc2vec (D2V) to learn paragraph vector embeddings that represent texts of arbitrary length (e.g., sentences, paragraphs, documents) in dense low-dimensional space. D2V is trained similarly to MP2V; likewise, two variants of D2V exist: distributed bag of words (DBOW) and a distributed memory model (DMM). The DBOW variant functions analogously to MP2V, with the model’s goal being to maximize the probability of document $D$ being constructed of the different words in the corpus’s vocabulary. DMM, by contrast, uses documents’ paragraph vector in addition to the ordered words in documents to sequentially predict words that appear next in the document. The authors note that this leads DMM to embed the topic of a document by representing what information is missing from it. Whereas the authors find that DMM generally perform better on prediction and classification tasks, others find that DBOW produces more accurate results (Lau and Baldwin 2016). For this reason, I will follow Le and Mikolov’s (2014) advice and also concatenate the embeddings separately generated by DBOW and DMM for downstream prediction tasks.

I extracted the submission text for all submissions in the main SW sample to train my D2V models. This resulted in 1,200,579 documents for training after submissions with no text were excluded (e.g., submissions
that posted a photo with no commentary). To preprocess documents for training, I used the Stanford CoreNLP toolkit (Manning et al 2014) and basic Python string methods to lowercase, convert non-ascii characters to similar ascii characters, recode digits as `<num>` and URLs as `<www>`, split slash-connected words like ‘black/white’ into ‘black’ and ‘white’, and finally tokenized all text by words. I then used gensim’s (Rehurek 2010) implementation of D2V to convert submissions’ text into a TaggedDocument object, where each submission is represented as a list of all tokenized words plus document tags for author and subreddit names. This step allows D2V to learn document embeddings for both authors and subreddits. Next, I trained each submission is represented as a list of all tokenized words plus document tags for author and subreddit (component. 

ward, all references to the sample will refer to the full network made of all subsamples combined into one large generate less representative and more imprecise embeddings for such subsamples’ nodes. From this point for-

ences in their respective averages and standard deviations. This indicates the forest fire models that generated
the samples drew comparable samples of nodes in which authors from no sample were substantially more or
less well-connected to others compared to authors from any other sample. Had such differences existed, they
could have added exogenous bias to MP2V’s sampling and embedding procedures, as after the four subsam-

less well-connected to others compared to authors from any other sample. Had such differences existed, they
the samples drew comparable samples of nodes in which authors from no sample were substantially more or
less well-connected to others compared to authors from any other sample. Had such differences existed, they
could have added exogenous bias to MP2V’s sampling and embedding procedures, as after the four subsam-

4. Results

4.1. Descriptive statistics and embedding visualizations

Figure 2 shows the total degree distribution for authors’ connections (made by collapsing edges from authors
to subreddits to other authors into author-to-author edges) among the four subsamples comprising the full sample. Authors have similar levels of connection to other authors across each subsample with very minor dif-

ferences in their respective averages and standard deviations. This indicates the forest fire models that generated
the samples drew comparable samples of nodes in which authors from no sample were substantially more or
less well-connected to others compared to authors from any other sample. Had such differences existed, they
could have added exogenous bias to MP2V’s sampling and embedding procedures, as after the four subsam-

amples are combined, highly connected nodes from subsamples with higher total degree distributions would be
sampled by the MP2V processes more often than nodes from subsamples with lower distributions. This bias at
the very least would affect how often certain nodes and their neighbors are embedded together, which would
generate less representative and more imprecise embeddings for such subsamples’ nodes. From this point for-
ward, all references to the sample will refer to the full network made of all subsamples combined into one large component.

To visualize the positions author and submission nodes in two-dimensional space, UMAP (uniform man-
ifold approximation and projection (McInnes et al 2018)) was used to reduce the dimension of nodes’ 128-
dimensional embedding positions. Figure 3 shows the UMAP results. Overall, figure 3 looks similar to standard
social network analysis visualizations (e.g., spring force directed projection) but without edges being drawn
between nodes (see supplemental figure 1 to compare MP2V to a spring force directed projection of the
author-subreddit projected network and see supplemental figure 2 to compare the MP2V results to those from
MP2V++). Various clusters of authors and subreddits are readily visible in figure 3. For example, on the right,
SW is clustered with depression, anxiety, and self-harm subreddits. A group of bipolar subreddits exists in the
lower-left corner. The bottom-right corner has a large cluster of subreddits related to various mental health
topics. Finally, the center and top-left corner both have clusters of general subreddits unrelated to mental health
or self-help (e.g., AskReddit). Table 2 shows cosine similarities between SW and its ten nearest neighbors. All
results reflect what one would expect—SW is positioned near other mental health subreddits on topics that
are highly related to suicidality, including depression, anxiety, self-harm, and social and emotional support (MMFB = make me feel better).

Both variants of the D2V model extracted similar but not overlapping language information from submissions’ text. Table 3 shows DBOW and DMM’s similarities in more detail. Overall, there is much simi-
arity among the two language embedding models, but as reflected by subreddit ordering arrows, DBOW and
DMM do represent different kinds of language similarities. For example, the top-15 nearest neighbors of SW in
DMM all have cosine similarities above 0.94, but DBOW similarities range from 0.96 to 0.89. Also, whereas depression is the nearest neighbor of SW among both models, other subreddits like MMFB are higher among the nearest neighbors in DBOW versus DMM. Many of the nearest neighbors in both D2V variants are also among the nearest neighbors to SW in the MP2V node embeddings, noted by subreddits underlined in bold. D2V and MP2V comparisons will be discussed more in section 4.3, however.

Table 4 presents examples of non-SW and SW submissions taken from submission authors in the main SW sample, plus DBOW and DMM cosine similarities to SW for submissions. The average and standard deviation
Figure 2. Total degree distributions for samples (via author-author edges). Notes: SW = SuicideWatch, MH = mental health, SH = self-help, R = random; author-to-author edges are measured by projecting the hybrid heterogeneous networks to one-mode networks. Total-degree $\bar{x}(\sigma)$: SW = 9.1 (0.63), MH = 9.0 (0.59), SH = 9.0 (0.56), R = 9.4 (0.64).

Figure 3. UMAP visualization of subreddit-author-subreddit metapath2vec embeddings. Note: node annotations include the following: author = orange, SuicideWatch = red, mental health subreddit = blue, self-help subreddit = purple, random subreddit = green, transparency = overlapping nodes. Text annotations include the following: SuicideWatch subreddit = red with border, SuicideWatch author = red, major mental health subreddit = blue with border, minor mental health subreddit = blue, self-help subreddit = purple, random subreddit = gray. All mental health and self-help subreddits are text annotated; however, only 5% of SuicideWatch authors (swauth labels in red) and 1% of random subreddits are annotated with text for their names.

of these DBOW and DMM similarity scores for the two sets of submissions (beyond what is shown on screen) are presented below their respective tables. A few qualitative observations on how DBOW and DMM models
Table 2. Cosine similarities between SuicideWatch and its 10 nearest neighbors.

|                      | Depression: 0.83 | Anxiety: 0.82 | Mentalhealth: 0.75 | AskDocs: 0.75 | MMFB: 0.74 |
|----------------------|------------------|---------------|--------------------|---------------|------------|
| Advice: 0.74         | Socialanxiety: 0.74 | Selfharm: 0.73 | Needafriend: 0.72   | StopSelfHarm: 0.72 |

Table 3. Cosine similarities to SuicideWatch among doc2vec models’ 15 nearest neighbors.

| DMM similarities to SuicideWatch: | DBOW similarities to SuicideWatch: |
|-----------------------------------|-----------------------------------|
| Depression, 0.99                  | Depression, 0.96                  |
| depressed, 0.98                    | MMFB, 0.93                        |
| Depression_help, 0.96              | Depression_help, 0.92             |
| Suicide_help, 0.97                 | Whatsbotheringyou, 0.92           |
| Getting_over_it, 0.96              | Depressed, 0.92                    |
| Prevent_Suicide, 0.96              | Suicide_help, 0.91                 |
| Mentalhealth, 0.96                 | sad, 0.91                         |
| MMFB, 0.95                        | Suicidenotes, 0.91                |
| SanctionedSuicide, 0.95            | offfmychest, 0.91                 |
| Venting, 0.95                      | SanctionedSuicide, 0.90           |
| Suicidenotes, 0.95                 | Getting_over_it, 0.90              |
| Mentalillness, 0.95                | venting, 0.90                     |
| ptsd, 0.94                         | Mentalhealth, 0.89                 |
| BPD, 0.94                          | Selfhelp, 0.89                     |

*Note: underlined subreddits in bold are also substantially close to SuicideWatch in MP2V embedding similarities.

Table 4. Examples of non-SuicideWatch and SuicideWatch submissions and their doc2vec cosine similarities to SuicideWatch’s subreddit-level doc2vec embedding position. Note: submission text has been altered so authors cannot be identified. Subreddit names in non-SuicideWatch subreddits are also adjusted to be similar to the theme of the original subreddit source.

| Non-SuicideWatch subreddit submissions | SuicideWatch submissions |
|---------------------------------------|--------------------------|
| Subreddit | Submission | DBOW Dist | DMM Dist | Subreddit | Submission | DBOW Dist | DMM Dist |
| video_game | Here is a guide to win the … | 0.26 | 0.17 | SuicideWatch | I have a friend who tried to … | 0.72 | 0.42 |
| need_advice | What really bothers me is … | 0.53 | 0.47 | SuicideWatch | I want to die and I don’t care if … | 0.74 | 0.50 |
| dating_tips | Rejected by the girl I love … | 0.71 | 0.42 | SuicideWatch | Keep having reoccurring thoughts … | 0.78 | 0.59 |

Table 5. Average (σ) doc2vec cosine similarity scores to SuicideWatch among submissions to different sets of subreddits.

|                      | Anxiety: DBOV 0.45 (0.15) | DMM 0.40 (0.09) |
|----------------------|---------------------------|-----------------|
| Overall:             | DBOV 0.74 (0.07)          | DMM 0.49 (0.06) |

track different kinds of language information can be drawn from these submissions and their similarities. First, the dating_tips submission in the non-SuicideWatch table has a high DBOW similarity but low DMM similarity, possibly because it has words that are common in SW submissions (e.g., rejection from a loved one) but excludes suicidal topics. Similarly, the top-most submission in the SW table has a high DBOW similarity but low DMM similarity, possibly because the topic is on the author’s friend. Finally, reoccurring thoughts is a strong indicator of suicidality, which may explain why the bottom-most submission in the SW table scores high for both DBOW and DMM. Table 5 shows the average and standard deviation of DBOW and DMM cosine similarity scores between submissions to other subreddits compared to SW. Overall, across all subreddits in the SW sample, submissions have relatively low cosine similarity to SW, yet submissions to anxiety, depression, and SW have increasingly similar submissions on average to SW, which indicates the two D2V models do distinguish between words and topics that are common in SW versus other subreddits.
In table 6, I show the Pearson correlation matrix of embedding similarities to SW. Overall, DBOW and DMM embeddings have a moderately strong correlation with each other; however, they are not perfectly overlapping—quantitatively indicating how the two D2V variants learn to represent different kinds of language information in submissions. The MP2V similarities are only slightly correlated with the two D2V similarities, which bodes well for avoiding multicollinearity in the prediction models while also giving evidence that the two forms of embedding models are tracking different kinds information about individuals’ behavior and communication and are not extracting redundant similarity data. Figure 4 shows scatterplots and histograms for the distributions between these similarities to visualize the shape of these correlations. All results are either circular or oval, indicating that heteroskedasticity will not bias the prediction models.

4.3. Prediction tasks
For the prediction tasks, I first truncated training/testing data to 305 K by filtering on submission authors from the main SW sample and then filtered the training/testing data to 293.5 K after any authors who did not have MP2V embeddings were dropped. To balance the training set, I randomly subsampled 4550 non-SW authors as negative non-SW cases to compare with 4060 SW authors. For the test set, I drew a random subset of 1015 SW

Table 6. Pearson correlation matrix of embedding similarities to SuicideWatch*.  

|           | MP2V | DBOW | DMM | D2Vx |
|-----------|------|------|-----|------|
| MP2V      | 1.00 | 0.23 | 0.15| 0.22 |
| DBOW      | 0.23 | 1.00 | 0.59| 0.93 |
| DMM       | 0.15 | 0.59 | 1.00| 0.83 |
| D2Vx      | 0.22 | 0.93 | 0.83| 1.00 |

*Note: D2Vx = the average of DBOW and DMM embedding similarities.

Figure 4. Scatterplots and histograms of metapath2vec and doc2vec embedding similarities to SuicideWatch.
authors and 1138 non-SW authors. I then used a logistic regression model with only the 128-dimension MP2V embeddings as features to predict whether an author will post in SW. Testing accuracy (69% ± 5%), the sum of true-positive and true-negative predictions divided by the total number of predictions across all prediction cells, was statistically greater than random guess (50% per prediction cell, given the balanced training); however, there were quite a few false-positives and false-negatives (25% and 38% respectively). Overall, however, this model performed quite well for having only included completely unsupervised positional data based on network behavior (figure 5–7).

Next, I ran the same prediction model for whether or not a user would post in SW using the same training/testing split as above but with only the two D2V models’ SW distances (i.e., with only 2 covariates instead of 128). Testing accuracy was still quite good at 76% (±5%). The predictive language embedding model
generated slightly fewer false-positives (21%) than the MP2V-only prediction model but still produced many false-negatives (27%). Overall, however, the model did remarkably well given that it only used two simple embedding distance measures.

Finally, to test how well one can predict if a user will post in SW, I integrated MP2V embeddings and both sets of D2V embedding distances. In addition, I tested this prediction model against all remaining 283 K observations without subsampling for the test set to assess how well the model could perform across the full sample (and thus how it may generalize to its analogous population). Subsampling was still used for training to balance the model’s learning, however, and so the training set had 4073 SW authors and 5792 non-SW authors while the test set had 1002 SW authors and 282 K non-SW authors. When used together, integrated MP2V and D2V embeddings predict with significantly higher accuracy (90% ± 3%) than the previous models and have high precision (89%) and recall (88%)—generating relatively few false-positives and false-negatives (10% and 12%, respectively). These large decreases in the false-positive and false-negative rates compared to the graph- and language-only models are noteworthy, as models with many more negative cases than positive cases can become biased to learn to predict ‘negative’ for each case, as doing so will be guaranteed to generate a high accuracy for the model. For example, in this final model, 99.6% of cases were truly negative, and so the model could have achieved 99.6% accuracy by simply learning to predict ‘negative’ for every case. This would lead the model to be wrong 100% of the time when predicting positive cases, however, which would be a terrible model for its intended use (predicting rare positive events). Instead, the balanced training helped the model successfully avoid this bias and generate results that are highly accurate as well as precise.

The improved performance of the integrated MP2V + D2V model makes intuitive sense, given the previous results I discussed relating to similarity differences between D2V’s DBOW and DMM models and how both variants produced different sets of nearest neighbors. Together, they may be learning to represent topics people are discussing, how people are talking about those topics, and where in the network they talk about different topics. In other words, authors’ behavior and their language are both important for predicting if authors will post in a particular subreddit—especially for reducing false-positives and false-negatives when target outcomes are rare in a dataset.

### 4.4. Sensitivity analysis

To qualitatively evaluate how well the integrated prediction model was able to predict if authors would post in SW, I used UMAP to generate a two-dimensional projection of the MP2V embeddings of authors and subreddits and annotated accurately predicted SW and non-SW authors. Figure 8 shows these results. Even when their MP2V embedding positions were located quite some distance away from SW and when they were embedded close to other general subreddits like AskReddit in the upper-left corner of the graph, the integrated model was
A methodological contribution that integrated graph and language embedding models could make is in analyzing the presentation of psychiatric attributes in text (e.g., suicidality, depression, anxiety, substance abuse, etc.) concerning the politician’s message, stance, or political party).

Not reflected in network ties alone (e.g., a Twitter user who replies to a politician’s tweet but does so to critique). For example, Qiu et al. (2018) showed how MP2V and D2V make significantly better than random predictions when used separately; however, the integrated prediction model using MP2V and D2V together generated highly accurate results with relatively few false-positive and false-negative predictions.

The results of this study extend research on the importance of simultaneously analyzing behavior and language in massive networks and efforts to integrate different kinds of embedding models for prediction and classification tasks, especially when they involve rare events. This is particularly important to emphasize given an ensemble of models—including deep neural networks, boosting, bagging, and stacking—only improved the integrated predictive model’s accuracy to 93%, which is not statistically different from the integrated model’s performance. The fact that the ensemble improved the integrated model’s performance so little can also be seen as evidence for how embedding models produce high quality dense low-dimensional representations of the large and complex sparse dataset they learn to embed. Though other papers have argued that simply concatenating embeddings from different models may lead to suboptimal results (Zhang et al. 2017), the computational complexity savings by this simpler model may make the tradeoff worth the difference. Given the accuracy of results produced in this context of predicting rare events, the tradeoff seems worthwhile.

Moving forward, fruitful avenues of research following from this work include using the integrated model to better test when and where one form of embedding helps improve performance more than using only one form of embedding model, understanding differences in membership among similar subreddits (e.g., comparing alcoholicsanonymous, AlAnon, cripplingalcoholism, stopdrinking, and addiction), and finally to predict the presentation of psychiatric attributes in text (e.g., suicidality, depression, anxiety, substance abuse, etc.). A methodological contribution that integrated graph and language embedding models could make is in analyzing and predicting social influence, social contagion, and other social dynamics. For example, Qiu et al. (2018) use graph embeddings to test neighbors’ social influence on individuals’ decisions to take different actions over time. Integrating language embeddings into this framework could help differentiate between individuals who share connections with others but who are different from them in cultural or ideological ways that are not reflected in network ties alone (e.g., a Twitter user who replies to a politician’s tweet but does so to critique the politician’s message, stance, or political party).

The application-specific results of this study also have important clinical and social relevance for mental health research, social support efforts, and interventions. First, I emphasized the importance of drawing representative network samples to improve models’ generalizability. Many studies of mental health dynamics on social media draw simple random samples of individuals; however, by foregoing a network-based sample, such studies lose the ability to directly incorporate information about how one’s network behavior, connectivity, and neighborhood effects may shape their results. Positive and negative social influence from supportive and potentially harmful network neighbors and network environments (e.g., hopeful versus despairing attitudes) are key predictors of whether individuals’ mental health will improve or decline, and network-based samples are able to extract this kind of structural and relational data whereas simple random samples extract
information on individuals as if they exist in isolation from one another. Studies testing social support can more accurately capture such dynamics through network-based samples.

Secondly, demonstrating that integrating graph and language data generates better predictions of suicidality aligns with clinical psychiatric diagnostic criteria which includes behavioral and verbal indicators of suicidality. These professionals use both forms of information to corroborate each other and avoid false-positive and false-negative diagnoses (e.g., in cases where one says they are not suicidal but has intentionally overdosed on medication). Lastly, predicting individuals at risk of becoming suicidal may allow for early interventions to help those at risk. Facebook is presently developing such systems in partnership with the crisis text line, the National Eating Disorder Association, and the National Suicide Prevention Lifeline. It is important to note, however, that such interventions may be detrimental to one’s self-esteem should they be inaccurate or unwanted (e.g., in cases where one is in fact receiving professional help and wishes to have privacy). Mental health is a sensitive topic, making those who struggle with it vulnerable. Any effort to implement system-wide screening and intervention systems, whether algorithmically- or user-driven, should be acceptable and welcomed by the given community. Otherwise, one may risk destroying trust and confidentiality in the group and cause people to leave it—defeating the purpose of social support communities and potentially causing further harm to individuals’ well-being.

While the results of this study are promising in many ways, they are qualified by a few limitations. First, neither the graph embeddings nor the language embeddings presented here are dynamically constructed. In other words, neither embedding learns to represent information on how individuals’ interaction or communication patterns evolve over time. Such data are important for predictions, as they can indicate whether one is moving toward or away from an outcome of interest. Some of the individuals predicted to be at high risk of suicidality in this study may have been at risk when they first joined Reddit but then moved farther from risk over time, whereas others may have had little risk when joining Reddit but then became increasingly at risk as time passed. Second, being at risk of suicidality in this study is defined as posting a submission in SuicideWatch. Not everyone who is suicidal may post in SuicideWatch (e.g., selection biases), and an estimated 22% of users who do post in SuicideWatch have no significant evidence of suicidality when their submissions are evaluated by a physician specializing in psychiatry (Ruch et al. 2019). Some users may also post in SuicideWatch with throwaway or secondary accounts. For example, some of the users in this study had words related to ‘throwaway’ in their name. Attempts to avoid such users were made by avoiding sampling users who had fewer than 20 submissions over their lifetime, but this method does not avoid problems when someone uses a throwaway or secondary account many times. Overall, these accounts would limit attempts to sample the true users’ complete network of interaction and language events and could lead to such users being sampled twice (i.e., a primary and secondary account are both sampled for one user). Lastly, since online contexts have different norms, incentives, and constraints than offline contexts, the integrated model presented here may not generalize to other contexts. Reddit, for example, provides more anonymity than Facebook, which may affect how people present themselves and interact with others.

Despite these limitations, the success of integrated graph and language unsupervised embedding models’ at predicting whether users post in a suicide support group (a rare event occurring among ~1% of this study’s sample) is appealing as it begs the question of whether ‘x2vec’-style models can save lives? As mentioned above, some social media platforms are already exploring if they can develop predictive models to detect if their users may be experiencing a mental health crisis. Most models created for this task use sophisticated language models developed in partnership with psychological, social, and mental health experts; however, these models may not be better than the unsupervised models proposed here, given unsupervised models’ ability to learn a variety of subtle and diffuse behavioral and language signals that supervised models may miss. The caveat here is that unsupervised models require many more positive and negative data observations from a wide variety of contexts to learn such cues; however, it is increasingly becoming easier to get such data multiple data sources. But even with the best model—supervised or unsupervised—identifying if a person is at risk of a mental health crisis will not save or help improve their life alone; no, models for predicting such events may save lives, but not unless they are used along with well-resourced and ethical intervention systems that provide such users with the support and assistance they need.

Acknowledgments

This work was supported in part by NSF Grant SES-1756822 and NIH Grant R25HD079352. The views and conclusions presented in this study are those of the author and should not be interpreted as necessarily representing the official policies or endorsements—either expressed or implied—of NSF, NIH, the US Government, Cornell University, or Graphika Inc. The author wishes to thank Drs Jennifer Ruch, Michael Macy, David Mimno, Thomas Gilovich, Vladimir Barash, James Moody, and Christopher Bail for feedback and support on parts of this project. The author is also grateful for assistance and feedback from Seunghyun Kim, Lillyan Pan,
Hannah Lee, James Zou, Jeffrey Tsang, Bryan Min, Juliana Hong, and Cornell University's Social Dynamics Laboratory.

Data availability

The data that support the findings of this study are openly available at the following URL: https://files.pushshift.io/reddit/.

ORCID iDs

Alexander Ruch https://orcid.org/0000-0003-4351-010X

References

Ahmed A, Shervashidze N, Narayananamurthy S, Josifovski V and Alexander J S 2013 Distributed large-scale natural graph factorization WWWW pp 1–11
An B, Chen B, Han X and Sun J 2018 Accurate text-enhanced knowledge graph representation learning Proc. of NAACL-HLT pp 745–55
Bail C A 2016 Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media Proc. Natl. Acad. Sci. 113 11823–8
Benamar F, Inkpen D and Taboada M 2018 Introduction to the special issue on language in social media: exploiting discourse and other contextual information Comput. Linguist. 44 663–81
Bojanowski P, Grave E, Armand J and Mikolov T 2017 Enriching word vectors with subword information Trans. Assoc. Comput. Linguist. 5 135–46
Borgatti S P and Groser T J 2015 Structural equivalence: meaning and measures International Encyclopedia of the Social and Behavioral Sciences 2nd edn (Amsterdam: Elsevier) pp 621–25
Cai H, Zheng V W and Kevin C-C C 2017 A Comprehensive Survey of Graph Embedding: Problems, Techniques and Applications vol XX
De Choudhury M, Kiciman E, Dredze M, Coppersmith G and Kumar M 2017 Discovering shifts to suicidal ideation from mental health content in social media AAAI pp 1–13
Cook B L, Progovac A M, Chen P, Mullin B, Hou S and Baca-Garcia E 2016 Novel use of natural language processing (NLP) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in Madrid Comput. Math. Methods Med. 2016 1–8
Corrigan P 2004 How stigma interferes with mental health care Am. Psychol. 59 614–25
Corrigan P W and Watson A C 2002 Understanding the impact of stigma on people with mental illness World Psychiatry 1 16–20
Crisp A, Gelder M, Goddard E and Meltzer H 2005 Stigmatization of people with mental illnesses: a follow-up study within the changing minds campaign of the Royal College of Psychiatrists World Psychiatry 4 106–13
Crisp A H, Gelder M G, Rix S, Meltzer H I and Rowlands O J 2000 Stigmatisation of people with mental illnesses Br. J. Psychiatry 177 4–7
Dong Y, Nitesh V C and Swami A 2017 Metapath2vec: scalable representation learning for heterogeneous networks KDD pp 135–44
Fortunato S and Barthelemy M 2007 Resolution limit in community detection Proc. Natl. Acad. Sci. 104 36–41
Fortunato S and Hric D 2016 Community detection in networks: a user guide Phys. Rep. 659 1–44
Frank D, Huang Z and Chyan A 2012 Sampling a large network: how small can my sample be? available at (http://snap.stanford.edu/class/cs224w-2012/projects/cs224w-036-final.pdf)
Frank K A 1995 Identifying cohesive subgroups Soc. Network. 17 27–56
Frank K A 1996 Mapping interactions within and between cohesive subgroups Soc. Network. 18 93–119
Glotsis G, Oellrich A, Velupillai S, Liakata M, Tim J, Hubbard P, Dobson R J B and Dutta R 2017 Characterisation of mental health conditions in social media using informed deep learning Sci. Rep. 7 45141
Gonzalez-Bailon S, Kaltenbrunner A, Banchs R E and Gonzalez-Bailon S 2010 The structure of political discussion networks: a model for the analysis of online deliberation J. Inf. Technol. 25 230–43
Goyal P and Ferrara E 2017 Graph embedding techniques, applications, and performance: a survey IEEE Trans. on Pattern Analysis and Machine Intelligence
Grohe M 2020 word2vec, node2vec, graph2vec, X2vec: towards a theory of vector embeddings of structured data PODS vol 20 pp 1–16
Grover A and Leskovec J 2016 Node2vec: scalable feature learning for networks KDD pp 855–64
Gundlapalli A V et al 2013 Using natural language processing on the free text of clinical documents to screen for evidence of homelessness among US veterans AMIA Annual Symp. Proc. pp 537–46
Seitz H and Kristy 2016 Language as a measure of mental health available at (https://www.clsp.jhu.edu/wp-content/uploads/2016/06/Language-as-a-Measure-of-Mental-Health.pdf)
Kennebeck S and Bonin L 2019 Suicidal behavior in children and adolescents: epidemiology and risk factors UpToDate® (https://www.uptodate.com/contents/suicidal-behavior-in-children-and-adolescents-epidemiology-and-risk-factors) (Accessed: 29 August 2019)
Kurant M, Markopoulou A and Thiran P 2010 On the bias of BFS (breadth first search) 22nd Int. Teletraffic Congress (LTC.22) pp 1–8
Lau J H and Baldwin T 2016 An empirical evaluation of doc2vec with practical insights into document embedding generation Rep4NLP pp 78–86
Le Q and Mikolov T 2014 Distributed representations of sentences and documents Proc. of the 31st Int. Conf. on Machine Learning pp 1–9
Leskovec J and Faloutsos C 2006 Sampling from large graphs KDD pp 1–8
Link B G, Cullen F T, Struening E, Shrout P E and Dohrenwend B P 1989 A Modified Labeling Theory Approach to Mental Disorders: An Empirical Assessment vol 54
Manning C D, Surdeanu M, Bauer J, Finkel J, Bethard S J and McClosky D 2014 The Stanford coreNLP natural language processing toolkit Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations pp 55–60
McInnes L, Healy J and James M 2018 UMAP: uniform manifold approximation and projection for dimension reduction (arXiv:1802.03426)
Mikolov T, Chen K, Corrado G and Dean J 2013 Distributed representations of words and phrases and their compositionality NIPS pp 1–9
Moser C, Groeneveld P, Huysman M, Moser C, Groeneveld P and Huysman M 2013 Extending social network analysis with discourse analysis: combining relational with interpretive data Influa. Technol. Soc. Network. Anal. Min. 6 547–61
Ou M, Cai P, Pei J, Zhang Z and Zhu W 2016 Asymmetric transitivity preserving graph embedding KDD pp 1–10
Peixoto T P 2014 Hierarchical block structures and high-resolution model selection in large networks Phys. Rev. X 4 1–18
Perozzi B, Al-Rfou R and Skiena S 2014 DeepWalk: online learning of social representations KDD pp 1–10
Qiu J, Tang J, Ma H, Dong Y, Wang K and Tang J and Microsoft Research 2018 DeepInf: social influence prediction with deep learning KDD pp 1–10
Šechter R 2010 Software framework for topic modelling with large corpora Proc. of the LREC 2010 Workshop on New Challenges for NLP Frameworks pp 45–50
Ruch A, Kim S and Ruch J 2019 Not all support is the same: multilabel psychiatric classification and prediction of social support among members of a large online suicide support group (unpublished)
Schreiber J and Culpepper L 2019 Suicidal ideation and behavior in adults UpToDate® (https://www.uptodate.com/contents/suicidal-ideation-and-behavior-in-adults) (Accessed: 29 August 2019)
Shi Y and Macy M 2016 Measuring structural similarity in large online networks Soc. Sci. Res. 59 97–106
Tang J, Qu M, Wang M, Zhang M, Yan J and Mei Q 2015 LINE: large-scale information network embedding WWW pp 1067–77
Wang D, Cui P and Zhu W 2013 Structural deep network embedding KDD pp 1–10
Xiao H, Huang M, Meng L and Zhu X 2017 SSP: semantic space projection for knowledge graph embedding with text descriptions Proc. of the 31st AAAI Conf. on Artificial Intelligence pp 1–7
Yang C, Liu Z, Zhao D, Sun M and Edward Y C 2015 Network representation learning with rich text information IJCAI Int. Joint Conf. on Artificial Intelligence 2015-Janua(IJCAI) pp 2111–7
Zhang D, Yin J, Zhu X and Zhang C 2017 Homophily, structure, and content augmented network representation learning Proc.-IEEE Int. Conf. on Data Mining, ICDM pp 609–18