JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 02/06 pp103−156
Volume 2, Number 2, June 2021
DOI: 10.23919/JSC.2021.0011

From Symbols to Embeddings: A Tale of Two Representations in Computational Social Science

Huimin Chen, Cheng Yang, Xuanming Zhang, Zhiyuan Liu*, Maosong Sun, and Jianbin Jin*

Abstract: Computational Social Science (CSS), aiming at utilizing computational methods to address social science problems, is a recent emerging and fast-developing field. The study of CSS is data-driven and significantly benefits from the availability of online user-generated contents and social networks, which contain rich text and network data for investigation. However, these large-scale and multi-modal data also present researchers with a great challenge: how to represent data effectively to mine the meanings we want in CSS? To explore the answer, we give a thorough review of data representations in CSS for both text and network. Specifically, we summarize existing representations into two schemes, namely symbol-based and embedding-based representations, and introduce a series of typical methods for each scheme. Afterwards, we present the applications of the above representations based on the investigation of more than 400 research articles from 6 top venues involved with CSS. From the statistics of these applications, we unearth the strength of each kind of representations and discover the tendency that embedding-based representations are emerging and obtaining increasing attention over the last decade. Finally, we discuss several key challenges and open issues for future directions. This survey aims to provide a deeper understanding and more advisable applications of data representations for CSS researchers.

Key words: Computational Social Science (CSS); symbol-based representation; embedding-based representation; social network

1 Introduction

Computational Social Science (CSS) refers to the fields that utilize computational approaches to model, simulate, and analyze social phenomena. CSS has received widespread attention and undergone rapid development over the past decade[1, 2]. It now includes numerous subfields, such as computational sociology, computational politics, and computational communication.

CSS is a data-driven field that was born due to the accessibility and analyzability of massive amounts of data[3]. With the fast development of Internet technology and mobile devices, large-scale multi-modal data have been produced and digitally recorded, such as friendship and posts on online social networks, purchase behaviour on e-commerce websites, and movement trajectories on mobile devices. These data provide us with an opportunity to mine meanings in social science directly and comprehensively from data, which include discovering the social phenomenon, such as news framing and public opinion, explaining the phenomena, and finding the causal relations, etc.

In general, we can summarize the operational framework of CSS: from data to meanings, as shown in Fig. 1. Note that the operational process is different from

• Huimin Chen and Jianbin Jin are with the School of Journalism and Communication, Tsinghua University, Beijing 100084, China. E-mail: huimchen1994@gmail.com; jinjb@tsinghua.edu.cn.

• Cheng Yang is with the School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China. E-mail: albertyang33@gmail.com.

• Xuanming Zhang, Zhiyuan Liu, and Maosong Sun are with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. E-mail: billyzhang07@outlook.com; liuzy@mail.tsinghua.edu.cn; sms@mail.tsinghua.edu.cn.

∗ To whom correspondence should be addressed.

† Huimin Chen and Cheng Yang contribute equally to this paper.

Manuscript received: 2021-04-14; revised: 2021-06-26; accepted: 2021-07-02

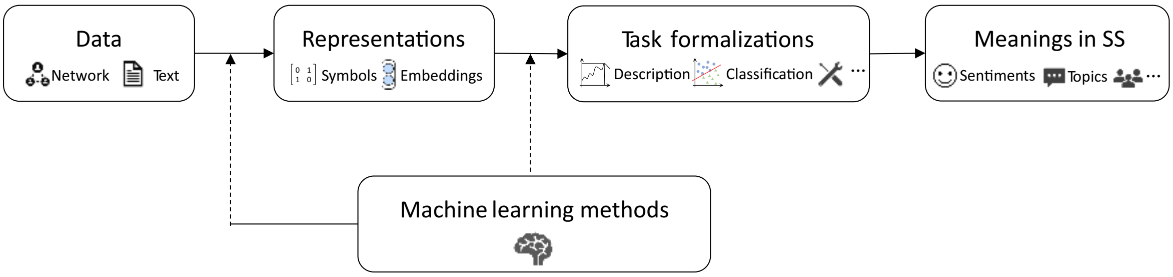


Fig. 1 Operational framework in CSS, where SS denotes social science.

the general research flow which can be problem-driven, followed by the selection of the required data, and then the identification of the task and the corresponding data representation. The operational framework we introduce here focuses on the implementation process of the study. Specifically, supposing we are conducting research in CSS, we first need to collect enough relevant data, which could be text or networks for our study. Afterwards, we need to transform the data into computationally processable representations, which are discrete or continuous numerals. Further, the representations of data are employed in practical applications, namely social issues we study. For each application, we formalize it into one of task prototypes, which commonly include data description, uncovering relationships between objects, clustering, classification, etc. Finally, the desired meaning in social science can be extracted based on the observation and analysis of the task results. Notably, the process from data to representations or representations to task formalizations usually requires the involvement of machine learning methods.

In the framework, we can find that the module of representations is not only the foundation, but also the key component since the increasing scale of data in CSS requires more efficient and effective representations. According to statistics in Ref. [4], there are now nearly 5 billion Internet users worldwide, who post hundreds of millions of tweets, view thousands of millions of videos on YouTube, and make billions of searches on the Google search engine every day. These massive data present us with a great challenge: how can we, the researchers in CSS, represent data effectively from such a large amount of multi-modal data?

Recently, the rapid development of data representation in computer science has nourished a large amount of successes both in academia and in industry[5]. Therefore, in this paper, we provide a systematic introduction for data representations that are divided into two schemes: symbol-based and embedding-based representations, as well as their existing applications in CSS to explore the effective and desirable data representations for different types of applications. We focus on the introduction of two most commonly used data, namely text and network, since they not only contain rich meanings but also are harder to represent, owing to the diverse expressions of text and complex structures of network.

To summarize, we make the following contributions in this survey:

- We provide a thorough review of data representations in two schemes: symbol-based and embedding-based representations, both for text and network. Researchers majoring in CSS can obtain a deep perception of these representations and distinguish them from each other clearly.
- We conduct a comprehensive survey on the applications utilizing the above representations, through investigating more than 400 top-cited articles from 6 representative publications over ten years. Based on the survey, we summarize the tasks in which each of the two representations excels, which can prompt the awareness of their expert areas and make advisable choices between them.
- We discover the trend that embedding-based representations are gaining growing attention, based on the statistics of their applications. This finding can encourage the usage of embedding-based representations in more relevant works and shed light on the future directions of CSS.

The rest of this survey is organized as follows. In Section 2, we present, in general terms, the definitions of symbol-based and embedding-based representations, and the differences between them. Afterwards, we meticulously introduce typical methods for constructing each kind of representations from text to network, in Sections 3– 6. In Section 7, we revisit the applications that use these representations and categorize them into different task prototypes in 6 top venues over past ten years. Based on the well-organized applications, we examine the coverage of the two representations and present their skilled areas in Section 8. In Section 9, we propose four open problems as well as future directions. Finally, we conclude the survey in Section 10.

2 A Tale of Two Representations

The representation indicated in this paper is behaved as computer-processable numerals, transformed from data in the real world. Each object (e.g., a word or a network node) in the real world can be assigned with a unique representation storing its characteristics. With the representation, we can conduct efficient analyses of large-scale data. It is the basis for data-driven CSS, since choosing an appropriate and exquisite representation will facilitate the subsequent analysis with fewer efforts.

Traditional representations are based on symbols. Following the definition from Wikipedia (https://en.wikipedia.org/wiki/Symbol), a symbol is " a mark, sign, or word that indicates, signifies, or is understood as representing an idea, object, or relationship". Hence, in this article, we identify symbol-based representations as discrete or continuous numerals which characterize objects in real-world explicitly and recognizably, such as language and relationship. It generally relies on the manual definition from data, which greatly contributes to the interpretability of CSS. For example, the representation of a word can be defined as its frequency in the corpus or sentiment value, while the representation of a node in the network can be designed as its degree or centrality.

Though symbol-based representation is explicit and human-readable, it suffers from several critical issues (Detailed issues of symbol-based representation are presented in Section 8). The most immediate shortcoming lies in heavy human efforts, since symbol-based representation is composed of manually defined features. To achieve a better performance, features should be elaborately designed. Besides, due to simple statistics and shallow combination of features, symbol-based representation usually fails to capture abstract semantics at a high level[5]. For example, humans can identify the similar semantic relation between "king"-"queen" and "man" - "woman" , while it is hard to discover for symbol-based representation.

To overcome these issues, the embedding-based representation is proposed to encode an object into a low-dimensional continuous vector, with the rapid development of artificial intelligence and deep learning methods. The vector is learned automatically by optimization of a training objective instead of hand-crafted features. It is randomly initialized and updated during the training process just like climbing a mountain step by step. Once the training is finished, we can use the learned embeddings as object representations for downstream tasks. Learning representations in such an automatic way is very convenient without human efforts. Moreover, it usually behaves as a complex combination of shallow features, which can detect the high-level attributes from data, such as the semantic relation mentioned above. But a shortcoming is that the interpretability of learned embeddings is poor, which means we usually have no idea about the exact meaning of embedding-based representations in each dimension.

In the following sections, we will introduce these two representations in detail, and further divide each kind of representation into text and network, namely symbol-based representations of text and network and embedding-based representations of text and network.

3 Symbol-Based Text Representation

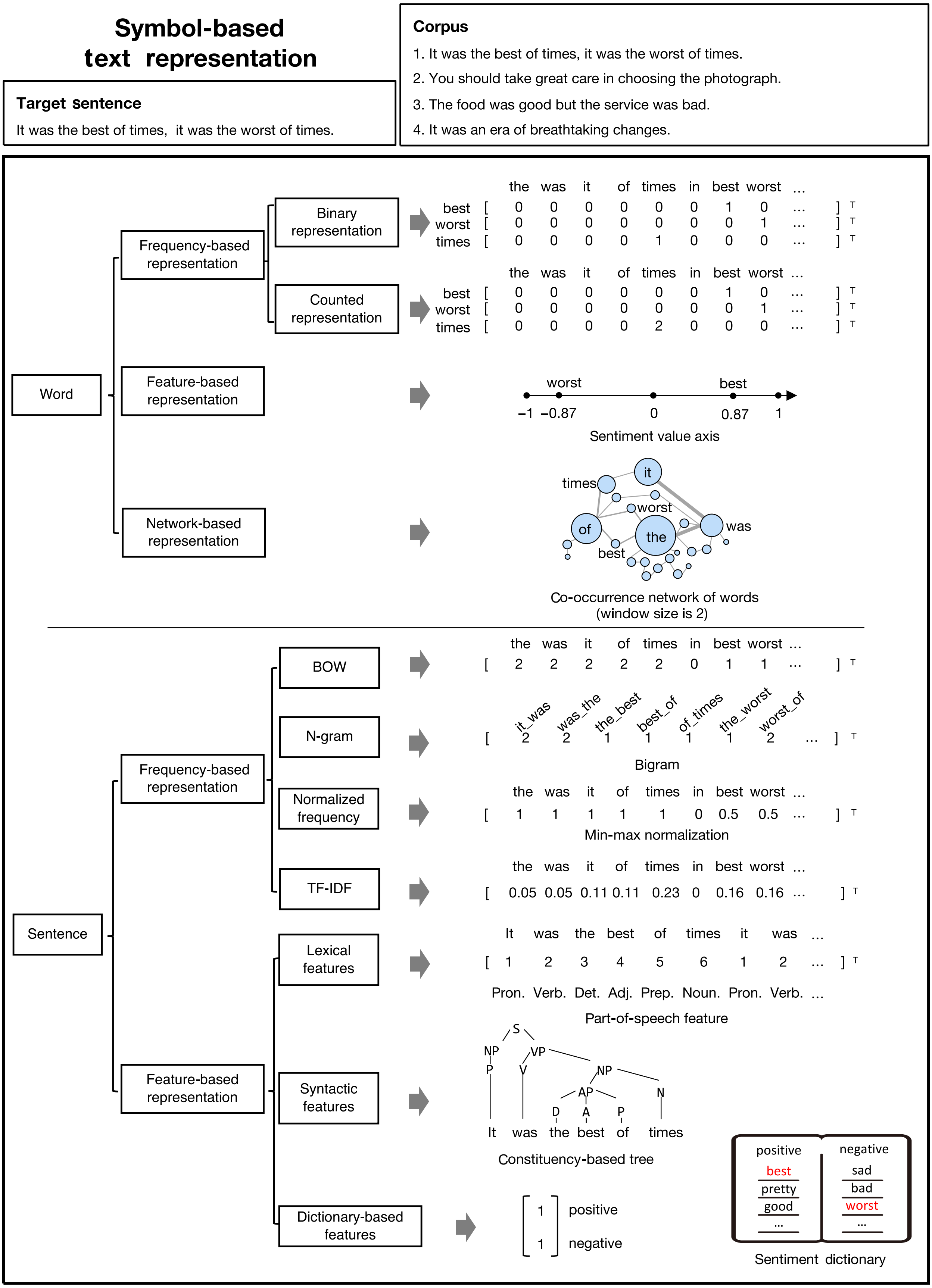
Text is the earliest and the most common form of data we use. In linguistics, a word is the smallest unit of text that can be uttered in isolation with objective or practical meaning. Phrases, sentences, and documents are all compositions of words. Therefore, in this section, we will first introduce the word representation which is the basis of representing texts. Afterwards, we will delineate the sentence representation based on symbols. Note that the representation of a document is similar to a sentence, since it can be viewed as a longer sentence or multiple sentences composed together. An illustration of symbol-based text representation is shown in Fig. 2.

3.1 Symbol-based word representation

Existing symbol-based word representations can be divided into three categories, namely frequency-based, feature-based, and network-based representations. In the following, we will introduce each of them in detail.

3.1.1 Frequency-based representation

Frequency is a basic statistic feature of words, reflecting



Symbol-based text representation

Target sentence

It was the best of times, it was the worst of times.

Corpus

1. It was the best of times, it was the worst of times.
2. You should take great care in choosing the photograph.
3. The food was good but the service was bad.
4. It was an era of breathtaking changes.

Word
Frequency-based representation
Binary representation
the was it of times in best worst ...
best [0 0 0 0 0 0 1 0 ...]ᵀ
worst [0 0 0 0 0 0 0 1 ...]ᵀ
times [0 0 0 0 1 0 0 0 ...]ᵀ
Counted representation
the was it of times in best worst ...
best [0 0 0 0 0 0 1 0 ...]ᵀ
worst [0 0 0 0 0 0 0 1 ...]ᵀ
times [0 0 0 0 2 0 0 0 ...]ᵀ
Feature-based representation
worst
best
−1 −0.87 0 0.87 1
Sentiment value axis
Network-based representation
times
it
worst
of
the
was
best
Co-occurrence network of words (window size is 2)
Sentence
Frequency-based representation
BOW
the was it of times in best worst ...
[2 2 2 2 2 0 1 1 ...]ᵀ
N-gram
it_was was_the the_best best_of of_times the_worst worst_of
[2 2 1 1 1 1 2 ...]ᵀ
Bigram
Normalized frequency
the was it of times in best worst ...
[1 1 1 1 1 0 0.5 0.5 ...]ᵀ
Min-max normalization
TF-IDF
the was it of times in best worst ...
[0.05 0.05 0.11 0.11 0.23 0 0.16 0.16 ...]ᵀ
Feature-based representation
Lexical features
It was the best of times it was ...
[1 2 3 4 5 6 1 2 ...]ᵀ
Pron. Verb. Det. Adj. Prep. Noun. Pron. Verb. ...
Part-of-speech feature
Syntactic features
S
NP
VP
P
V
NP
AP
N
D
A
P
It was the best of times
Constituency-based tree
Dictionary-based features
1 positive
1 negative
positive
best
pretty
good
...
negative
sad
bad
worst
...
Sentiment dictionary

Fig. 2 An illustration of symbol-based text representation.

the significance of words in the corpus. Frequency-based word representation transfers each word into a value or a vector based on its occurrence in the corpus. Specifically, it can be categorized into two settings:

Binary representation. Each word is denoted as $0/1$ depending on whether it appears in the corpus or not. Taking the word "times" in the target sentence in Fig. 2 as an example, it is represented with value 1 as it appears in the corpus, while "time" is represented with value 0 due to absence. Further, each word can also be indicated as a vector with its dimension size equal to the vocabulary size, i.e., the number of all words in the corpus. Each word is assigned with a unique index at first, then its vector behaves as that all elements are zeros except the only dimension of its index is one. As shown in Fig. 2, "times" is represented as a vector $[0,0,0,0,1,0,\dots]$, only the dimension indicating itself is 1. Hence, it is also known as *one-hot representation*, with its dimension probably being tremendous if given large vocabulary size.

Counted representation. Distinguished from binary representation, each word is expressed based on its number of occurrences in the corpus. For example, we can denote "times" as its count: 2, or a vector with the value in the dimension of its index to be the count: $[0,0,0,0,2,\dots]$. These two types of representations are corresponding to the value and vector in the binary representation, respectively. The difference is that they introduce information of the word's occurring number in this counted representation fashion.

3.1.2 Feature-based representation

Apart from the frequency-based approach, feature-based representation signifies each word with manual features defined depending on the research goal. For example, a word can be represented as a vector composed of its occurrences with designated words when measuring its semantics in some specific aspects. It also can be denoted as a human-defined sentiment value when considering its sentiment feature, as shown in Fig. 2. The "worst" is assigned to a sentiment value close to -1, while the "best" is arranged to be nearly 1.

3.1.3 Network-based representation

Substituting for representing each word directly as a value or a vector, network-based representation maps each word into a node in the network, where each edge between two nodes is established based on defined relations, such as occurrences or semantic relations. In the light of the constructed network, we can represent each word with its degree, centrality, closeness, and neighboring nodes, etc. As shown in Fig. 2, each word in our example corpus is projected into a node in the word co-occurrence network. This representation manner allows for better modeling of the characteristics of words and the complex relationships between words through utilizing a range of network analysis algorithms.

3.2 Symbol-based sentence representation

Symbol-based sentence representation is usually built upon word representation and can be separated into two groups: frequency-based representation and feature-based representation. In this section, we will describe them in detail.

3.2.1 Frequency-based representation

Frequency-based representation of sentences is constructed upon raw frequencies of words and phrases, as well as processed frequencies. As for the way based on raw frequencies, Bag-of-Words (BOW)[6] and n-gram representations[7] are widely utilized, while representation of normalized frequency and Term Frequency-Inverse Document Frequency (TF-IDF)[8] are commonly used with regard to the way based on processed frequencies. Below we will present each of them, using the target sentence in Fig. 2 as an illustrative example.

BOW. A sentence is represented as the bag of its words, where word order and grammar are disregarded, and only word frequency is kept. For example, the bag of target sentence in Fig. 2 contains 2 "the", 2 "was", and some other words with different numbers, which compose the representation vector of the target sentence. We can see that it is a simple representation, which is the sum of the one-hot representation of each word in the sentence.

***n*-gram.** Note that the word order information is disregarded in the BOW representation, resulting in that the two sentences "Good, not bad" and "Bad, not good" will have the same representation, though they have completely different semantic meanings. Therefore, n-gram (i.e., n consecutive words in a given sentence) count instead of word count is proposed. "not bad" and "not good" are two distinct 2-grams (or bigrams), so that the semantics of above sentences can be distinguished through 2-gram representations.

Normalized frequency. Replacing raw frequency, each sentence is performed as a normalized version of its raw frequency. Two of the most prevailing normalization methods are Min-Max and Z-score normalization, with the first mapping the value of raw frequency into the range of $[0,1]$ and the latter transferring data into a standard normal distribution. Through this manner, representations of all sentences can be transformed into the same order of magnitude, enabling the comparison between sentences in different magnitudes, such as measuring semantics similarity between sentences with quite different lengths. It can benefit the efficient execution of downstream tasks as well.

TF-IDF. Frequencies of word and n-gram are the only considered features in the above representations. However, we can see that words with the most frequencies are not always the most important. For instance, "a", "an", and "the" are all frequent words but usually without substantial meaning. Therefore, TF-IDF representation is proposed to further consider the document frequency, which is inspired by that a term's importance will decrease with the number of documents where it appears. Specifically, each value in BOW or n-gram representation is replaced with

$$\mathrm{tf}-\mathrm{idf}(w,d)=\mathrm{tf}(w,d)\times\mathrm{idf}(w,D) \tag{1}$$

where $\mathrm{tf}(w,d)$ denotes frequency of the term w in document d and $\mathrm{idf}(w,D)$ is inverse document frequency of the term w in corpus D. It keeps a balance between the term frequency in a sentence and the document frequency of the term in a corpus. Through the processing of raw frequency, TF-IDF representation can re-weight words and catch the important ones of a sentence or a document.

3.2.2 Feature-based representation

Feature-based representation is the most commonly utilized symbol-based sentence representation. It relies on artificially defined features, which can be divided into three main categories: lexical features, syntactic features, and dictionary-based features. The first two are based on features extracted from the text itself, and the last one depends on external dictionaries to obtain features. We then describe them in detail.

Lexical features. Specific words are distilled from the text as features, such as adjectives, adverbs, emoticons, and hashtags, which are informative lexicon features for downstream tasks. For example, adjectives and emoticons are central features for psychological studies, and verbs and nouns are particularly important when we intend to unearth topics from text.

Syntactic features. Each sentence is equipped with a specific syntactic structure, which also plays a crucial role in the semantics of the sentence. For example, as for the sentence "freedom is dearer than life", its syntactic structure can inform that "freedom" is the nominal subject of "dearer" rather than "life", providing the key information of semantics. Hence, syntactic features of the sentence are prevailingly extracted, using syntactic analysis (i.e., parsing) or manual designed rules, such as polarity shifts due to connectors and negations. Syntactic analysis is generally divided into constituency parsing and dependency parsing, with the former concentrating on breaking sentences into sub-components, such as sub-phrases, and the latter focusing on word connections based on their grammatical relations. Constituency parsing of part of the target sentence is shown in Fig. 2.

Dictionary-based features. Different from the above two kinds of features, dictionary-based features are recognized in the light of human-constructed dictionaries, such as Linguistic Inquiry and Word Count (LIWC)[9] and Language assessment by Mechanical Turk (labMT)[10]. Among these dictionaries, LIWC is the most widely adopted, where each word falls into several pre-defined dimensions, such as linguistic (e.g., person pronouns and conjunctions), psychological (e.g., anger and anxiety), cognitive dimension (e.g., insight and causation). It is worth mentioning that the difference from the lexical features lies in that dictionary-based features assimilate knowledge and wisdom summarized and accumulated in previous studies. Supposing there are two dimensions of words in a dictionary, i.e., positive and negative words, the target sentence can be represented in a two-dimensional vector $[1,1]$, with the first dimension indicating one positive word "best" and the second denoting one negative word "worst" occurring in the sentence, as shown in Fig. 2.

4 Symbol-Based Network Representation

A network (or graph) contains a set of objects and their relationships. An object is usually represented by a node (or vertex), and the relationship between two objects is represented by an edge between corresponding nodes. An edge can be directed to indicate an asymmetric relationship, weighted to emphasize the strength of a relationship, signed to represent a relationship is positive or negative, and etc. Most work will use adjacency list or adjacency matrix as the basic representations of a network. Then they will employ statistics or specialized modeling to build high-level representations.

4.1 Basic representations

Now we will start by presenting two basic representations of networks.

4.1.1 Adjacency list

Adjacency list is a collection of unordered lists where each list describes the set of neighbors of a node in the network. Taking the triangle structure in Fig. 3 as an example, the corresponding adjacency list contains three lists: $a:\{b,c\}$, $b:\{a,c\}$, and $c:\{a,b\}$. The adjacency list representation can record all edges in a space-efficient manner and be suitable to describe an (un)directed graph structure.

4.1.2 Adjacency matrix

Adjacency matrix is a square matrix whose dimension equals to the number of vertices. Each element of the adjacency matrix indicates a directed edge between the corresponding nodes. The adjacency matrix representation of Fig. 3 is

$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}.$$

The adjacency matrix representation can be used to describe (un)directed/weighted/signed graph structures by changing the ones to real-valued weights or signs. We can efficiently check whether two specific nodes are connected using the adjacency matrix representation. However, real-world networks are usually sparse, which means most elements in an adjacency matrix are zeros. The storage usage of an adjacency matrix is proportional to the square of the number of vertices, which is not space-efficient compared with the adjacency list representation.

4.2 Statistics on a network

The aforementioned adjacency list and matrix can faithfully record the structure of a network. However, in many scenarios, we need to extract features from a network, e.g., by statistics. We classify the statistics on a network into node/edge-based statistics and subgraph-based statistics.

4.2.1 Node/edge-based statistics

Note that node/edge-based statistics are not necessarily used to represent a node or an edge. For example, node degree can be used to represent a node, while average degree characterizes the entire network. To characterize and represent a network (or subgraph), we can calculate the size of a network (the number of nodes and edges), average degree, edge density (the ratio of the number of edges to the number of possible edges), etc. In fact, such statistics are widely used to describe the datasets.

In general, employing statistics to represent nodes is more common and useful in the studies of CSS because they usually need to model the behaviours or properties of individuals in a large (social) network. On one hand, the simplest statistics directly come from a node's behaviours or features, e.g., the number of a Facebook user (node)'s posts. On the other hand, the statistic-based representation can also come from a node's neighborhood structure. We will take local cluster coefficient as an illustrative example: As shown in Fig. 4, the local clustering coefficient of a node identifies the local density, and is defined by the proportion of the number of links between its neighbors divided by the number of links that could possibly exist between them. In addition, the statistics are also possible to be a mixture of node behaviours and network structure, e.g., the number of likes obtained from one's friends in an online social network.

4.2.2 Subgraph-based statistics

Subgraph-based statistics can be further categorized into cluster-based and motif-based.

A cluster in a network contains a group of nodes with dense connections or similar characteristics. Clusters in a network can be either overlapped or disjoint. A cluster is also referred to as a community in many scenarios. The cluster assignment of a node can be used as its cluster-

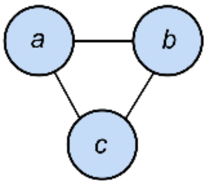


Fig. 3 An example network for illustration.

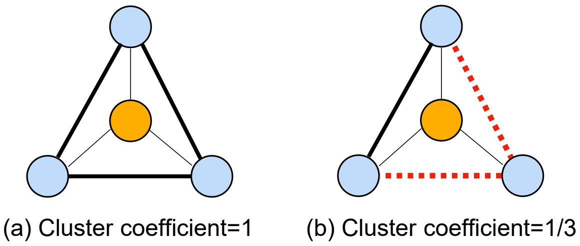


Fig. 4 An example of cluster coefficient.

based representation, as shown in Fig. 5. Besides, cluster-based indices can be used to characterize the whole network as well. For example, modularity measures the strength that a network is divided into clusters: the fraction of the edges within the clusters minus the expected fraction if edges are randomly distributed. A larger modularity indicates dense connections within clusters and sparse connections between different clusters.

On the other hand, motifs, which are defined as recurrent and statistically significant subgraphs or patterns, are much smaller than communities, e.g., a triangle made up of 3 nodes or a square made up of 4 nodes. The frequencies of motifs are widely used as motif-based statistics. As shown in Fig. 5, we can count the numbers of appearances of triangles and squares to

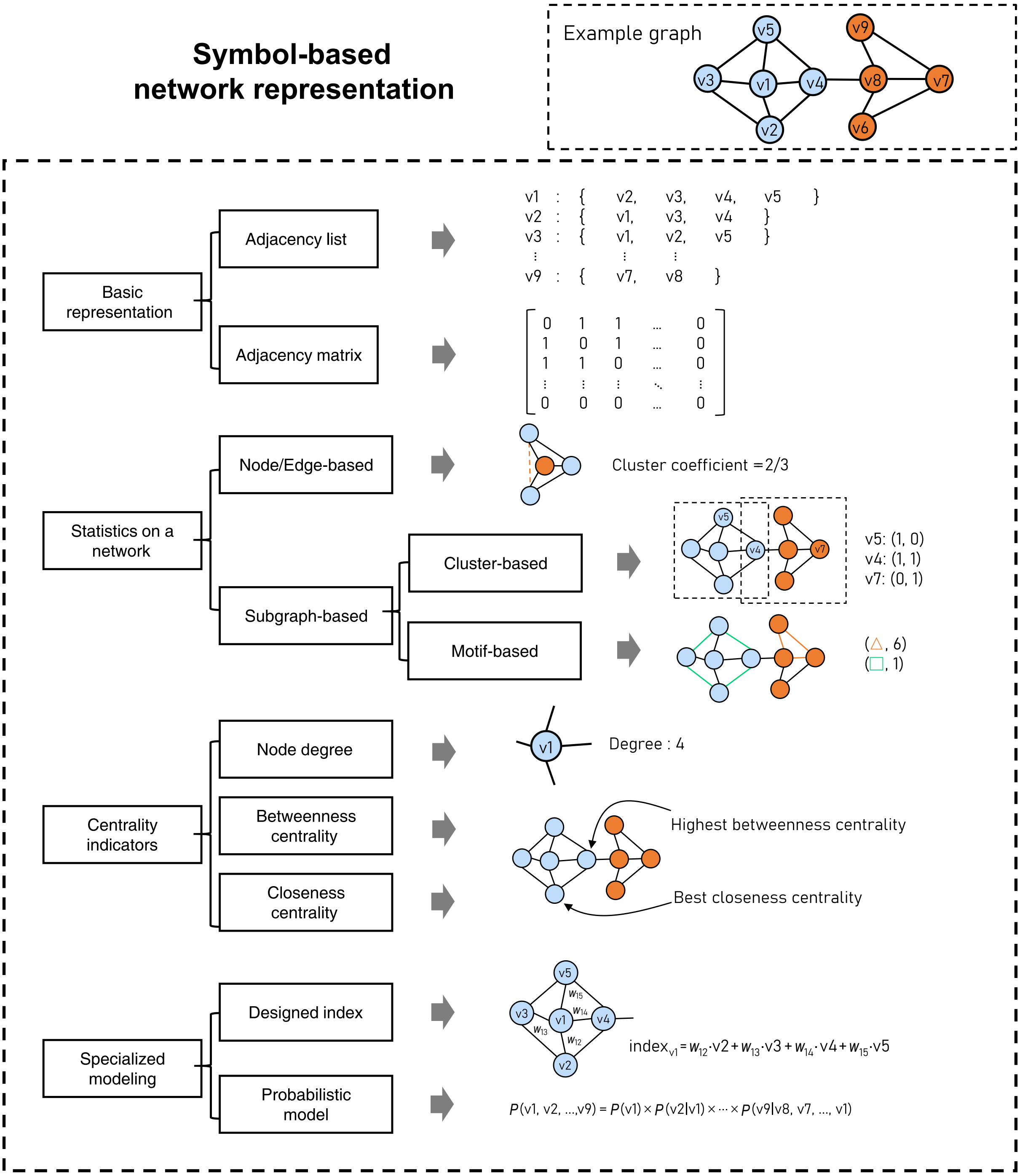


Fig. 5 An illustration of symbol-based network representation.

represent the entire network. In addition, the global cluster coefficient, which is calculated as the proportion of the number of closed triplets (i.e., triangles) divided by the number of all triplets (either closed or not), can give an indication of the clustering in the whole network.

4.3 Centrality indicators

To characterize the properties of nodes in a network, there exist various of indicators ranging from simple statistics to designed indices. Among all such pre-defined or manually designed indicators, centrality indicators, which measure the importance of each node in a network, are the most widely used ones and thus we put them into a separate subsection.

Node degree, i.e., the number of edges connected to a node, is the simplest centrality indicator. Intuitively, a node with a larger degree will have a larger impact on the network. Besides, *closeness centrality* of a node measures the average length of the shortest path between the node and all other nodes in the network. Hence, a node with smaller closeness centrality will be closer to all other nodes and thus be more central. *Betweenness centrality* counts the number of times a node acts as a bridge along the shortest path between two other nodes. A node with larger betweenness centrality will probably control the information flow or communications in the network. Figure 6 shows the nodes with best degree/closeness/betweenness centrality. There are also many other centrality indicators, such as eigenvector and PageRank, and readers are encouraged to learn more about them if interested (https://en.wikipedia.org/wiki/Centrality).

4.4 Specialized modeling

Real-world interaction systems are quite sophisticated and thus motivate many case-by-case representations of networks. Depending on how complicated a network representation is, we roughly divide them into designed index and probabilistic model.

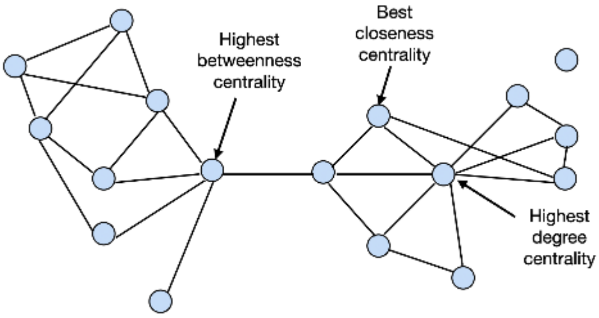

Highest betweenness centrality
Best closeness centrality
Highest degree centrality

Fig. 6 An example of centrality indicators.

Designed indices are usually a heuristic combination of multiple simple factors. For example, if we want to quantify how good a person works in a collaboration network, we can compute the weighted sum of his/her scores of error rate, decision time, and peer evaluation. In detail, the score of decision time could be an exponentially time-decayed function. In contrast, probabilistic models are much more complicated. Besides the probabilistic modeling among a number of variables, differential equations are also widely used to characterize the dynamics in a network. In all, both designed index and probabilistic model are usually more complicated than previously mentioned simple statistics and highly specialized for a given problem.

5 Embedding-Based Text Representation

Since text consists of multi-grained units as mentioned in Section 3: from words to sentences, embedding-based text representation also follows the same composition principle. In this section, we will introduce the most widely used method to learn the embedding-based representation of words and sentences, respectively. An illustrative demonstration is shown in Fig. 7.

5.1 Embedding-based word representation

Approaches of learning embedding-based word representation aim to embed each word into a low-dimensional and dense vector, and require that closer distance between two vectors in the space denotes more similar semantics between the corresponding words. The intuition behind these approaches is simple: words sharing similar contexts should have similar word embeddings. For instance, the word "apple" and "banana" will probably both appear in the context "I like eating xxx" or "xxx trees" from a large corpus, and thus should have similar word vectors. Existing methods fall into two main groups, namely count-based models and prediction-based models[11]. Next, we present each of them, respectively.

5.1.1 Count-based models

Count-based models establish distributional representations of words upon co-occurrence counting. A primary branch of these models works on transforming the co-occurrence matrix of words into a reduced space, with matrix factorization techniques, such as singular value decomposition (e.g., Latent Semantic Analysis (LSA)[12]) or weighted least-squares

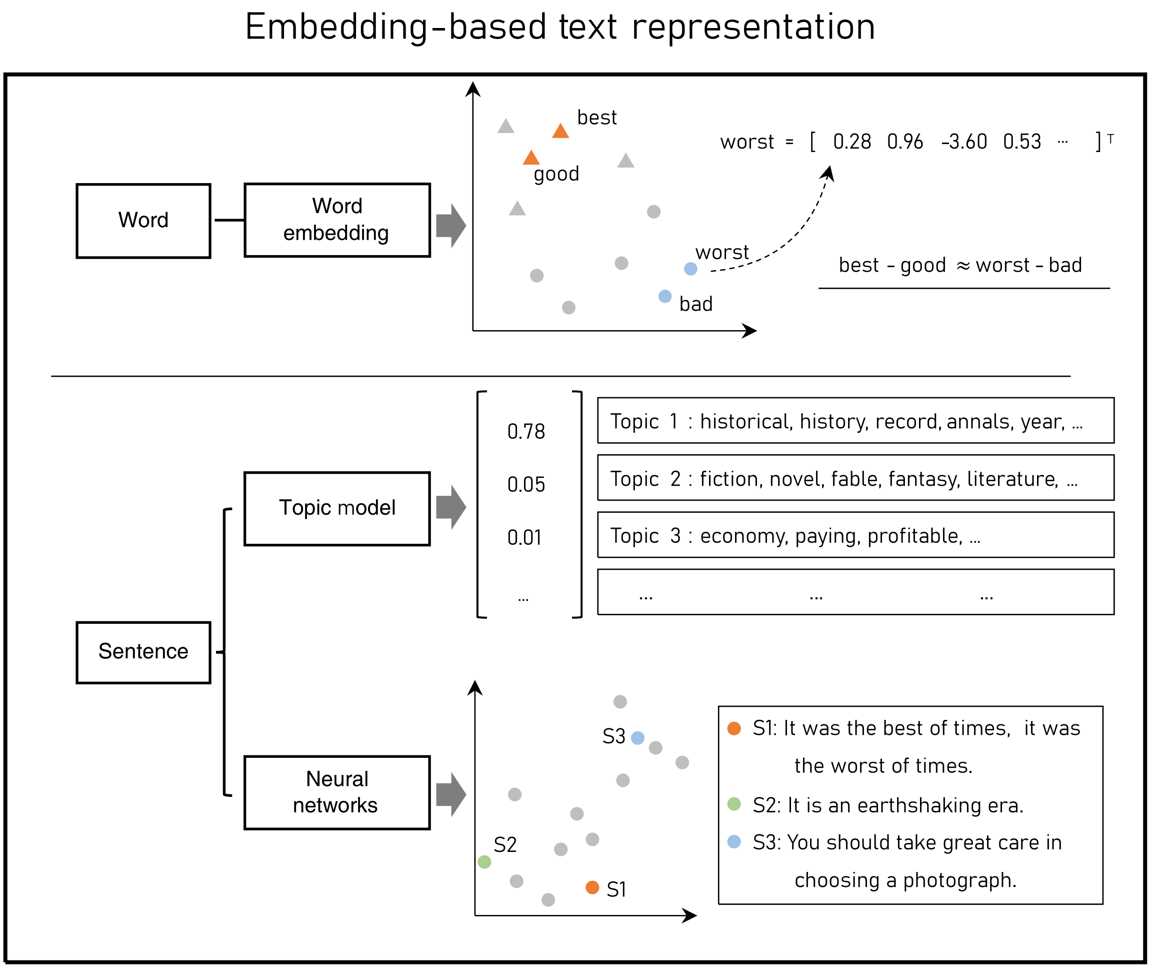

Embedding-based text representation
Word
Word embedding
best
good
worst
bad
worst = [0.28 0.96 -3.60 0.53 ···]ᵀ
best - good ≈ worst - bad
Sentence
Topic model
0.78
0.05
0.01
...
Topic 1 : historical, history, record, annals, year, ...
Topic 2 : fiction, novel, fable, fantasy, literature, ...
Topic 3 : economy, paying, profitable, ...
... ...
Neural networks
S1
S2
S3
S1: It was the best of times, it was the worst of times.
S2: It is an earthshaking era.
S3: You should take great care in choosing a photograph.

Fig. 7 An illustration of embedding-based representation of text.

regression (e.g., Global Vectors for word representation (GloVe)[13]). A brief example of LSA is shown in Fig. 8. Another branch of count-based models is Random Indexing (RI)[14], which learns distributional representation by assigning a randomly initialized vector to each word, and then gradually updating the vector according to the co-occurring contexts. It overcomes the difficulty of LSA by precluding expensive pre-processing of huge word-document matrices.

5.1.2 Prediction-based models

Prediction-based models aim to create low-dimensional distributional representations through optimization of the probability that predicts a target word based on contexts or predicts the contexts of a target word. Word2vec[15] is one of the most popular toolkits of prediction-based model proposed by Google in 2013, which can efficiently learn word embeddings from a large corpus. It is equipped with two model variants: Continuous Bag-Of-Words (CBOW) and Skip-Gram.

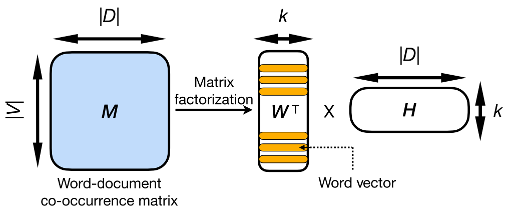


Fig. 8 An example of count-based model LSA.

CBOW optimizes a training objective of predicting a target word given its context words. As shown in Fig. 9, CBOW predicts the center word given a window of context with the window size l. The window size l is a hyper-parameter to be tuned.

Formally, CBOW predicts the probability of the i-th word w_i in the corpus, given its contexts of window size l,

$$\Pr(w_i|w_{i-l}\ldots w_{i-1}, w_{i+1}\ldots w_{i+l}) = \mathrm{softmax}(\boldsymbol{M}_c(\sum_{j:|j-i|l, j\neq i} \boldsymbol{w}_j)) \tag{2}$$

where softmax() is a normalization function that ensures the sum of the components of the output vector equals to

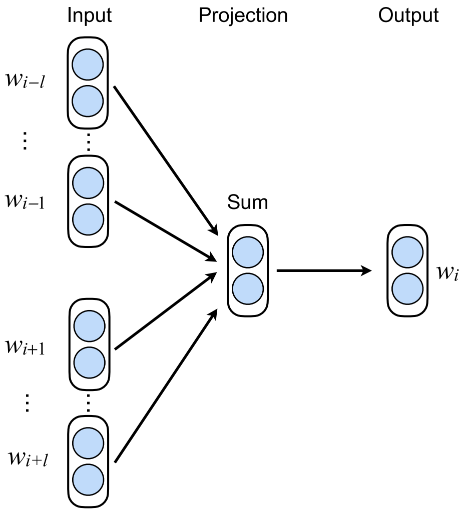


Fig. 9 Architecture of CBOW model.

1, $\boldsymbol{w}_i$ is the word vector of word w_i, $\boldsymbol{M}_c$ is the weight matrix in $\mathbf{R}^{|V|\times m}$, V indicates the vocabulary, and m is the dimension of word vectors. Then CBOW is optimized by maximizing the log likelihood,

$$\mathcal{L}_c = \sum_i \log \Pr(w_i | w_{i-l} \dots w_{i-1}, w_{i+1} \dots w_{i+l}) \quad (3)$$

Skip-Gram aims to predict the context words given a center one, as shown in Fig. 10. Formally, given a word w_i, Skip-Gram predicts each word $w_j (|j - i| \leqslant l, j \neq i)$ in its context,

$$\Pr(w_j | w_i) = \mathrm{softmax}(\boldsymbol{M}_s \boldsymbol{w}_i) \quad (4)$$

where $\boldsymbol{M}_s$ is the weight matrix. The optimization objective is defined as

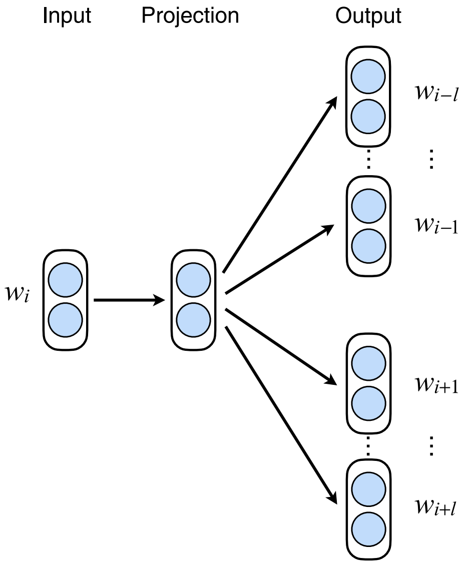


Fig. 10 Architecture of Skip-Gram model.

$$\mathcal{L}_s = \sum_i \sum_{j:|j-i|\leqslant l, j\neq i} P(w_j | w_i) \quad (5)$$

Word2vec further employs hierarchical softmax[16] and negative sampling[17] to speed up the computation process.

Though the algorithms differ, empirical results show that count-based models, such as GloVe, and prediction-based models, such as CBOW, perform comparably on semantic similarity and downstream tasks with certain system designs and optimized hyperparameters[18]. Hence, we uniformly refer to them as word embedding-based representations.

5.2 Embedding-based sentence representation

Similar to word embedding, embedding-based sentence representation is also formed as a continuous and dense vector with rich semantic meanings. There are two main series of methods to learn the sentence representation: one is based on topic models, another is based on neural network models. Below we present each of them in detail.

5.2.1 Topic model-based representation

Topic models seek to represent a sentence (document) as a distribution of a series of topics, based on two assumptions: each document contains multiple topics; each topic contains multiple words. Here, we describe the most typical topic models, including LSA, Latent Dirichlet Allocation (LDA)[19], and Structural Topic Model (STM)[20].

LSA is one of the basic techniques for topic modeling, of which the core idea is to decompose the document-word matrix into independent document-topic matrices and topic-word matrices. Then each row vector in the document-topic matrix can be used to represent the corresponding document. However, the meaning of each dimension (i.e., topic) in the row vector is vague to us, though we can measure the similarity between two documents by calculating cosine similarity between two row vectors.

LDA is the most widely used topic model and a member of the probabilistic graphical model. It introduces a probabilistic interpretation to the basic LSA through a generative model. Here we introduce the basic generative process of LDA, as shown in Fig. 11.

(1) For each document d, randomly choose a topic distribution θ_d over K topics from the prior Dirichlet distribution with hyperparameters α.

(2) For each word $w_{d,n}$ in the document,

- randomly sample a topic $z_{d,n}$ from the topic

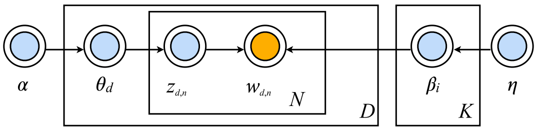


Fig. 11 Architecture of graphical model for LDA.

distribution θ_d;

- randomly choose a word distribution $\beta_{z_{d,n}}$ of topic $z_{d,n}$ over N words, from another prior Dirichlet distribution with hyperparameters η;
- randomly sample the word $w_{d,n}$ from the word distribution $\phi_{d,n}$.

Through this process, each document can be granted a representation (i.e., θ_d) denoting the distribution over topics, with each topic assigned a probability distribution (i.e., $\beta_{z_{d,n}}$) over words. With the help of topic-word distribution, we can further capture the keywords of each topic and elucidate the meaning of each topic.

STM further extends LDA to account for meta-data of text, since documents usually entail time, geographic location, author, title, and other additional information. These can be formalized as covariates in the topic model, so that each document can have its own prior distributions over topics and words depending on its covariates. This approach is widely used in CSS owing to the consideration of environmental variances of documents.

5.2.2 Neural-based representation

Neural-based representation is learned from neural network models, which are constructed based on a collection of connected artificial neurons inspired by the biological brain. These neurons are connected by edges with different weights which can be learned from the training process. The training process is operated by processing instances, each of which contains a given “input” and “output” . Once training begins, neural network models will update their weighted associations to bridge the gap between inputs and outputs. At the end of the training process, the sentence representation will be refined automatically without manual design.

Besides, the sentence representation based on neural network models can capture the complex internal structures of sentences owing to the flexible connections of neurons, such as sequential, hierarchical, and tree structures, which are essential for understanding sentences. Furthermore, neural network models allow us to imitate the cognitive mechanisms of the human brain, such as working memory[21] and attention mechanism[22], to construct sentence representation.

In the following, we will introduce the most popular used neural network models for learning embedding-based sentence representation, including Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Transformer.

CNN learns the sentence representation by two layers[23]: a convolution layer and a pooling layer, as shown in Fig. 12. The convolution layer extracts local features of the inputted sentence through multiple different filters. Formally, it behaves as a matrix multiplication between a convolution matrix and a sequence of word vectors in a sliding window centered on each word in the sentence. Afterwards, the pooling layer merges all local features to obtain a fixed-sized representation, with the max-pooling and mean-pooling layers most commonly used. These two layers can be represented as

$$\boldsymbol{h}_c = \text{Pooling}[f(\boldsymbol{W}_c \cdot \boldsymbol{x}_i + \boldsymbol{b}_c)] \tag{6}$$

where $\boldsymbol{W}_c$ denotes the convolution matrix, “Pooling” indicates the pooling layer, $\boldsymbol{x}_i$ denotes the concatenation of word representations in the subsequence centered on the i-th word, $f()$ and $\boldsymbol{b}_c$ indicate a non-linear function and a bias vector in the convolution layer, respectively, and $\boldsymbol{h}_c$ is the final sentence representation obtained from CNN model.

To sum up, CNN adopts the convolutional layer so that it can focus on the sentence’s local features and

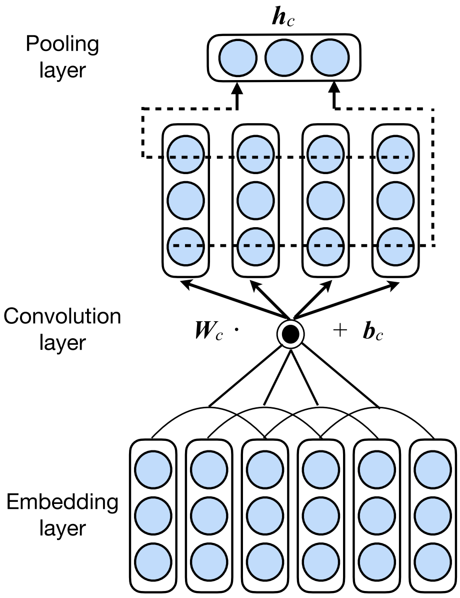


Fig. 12 Architecture of CNN.

effectively reduce the parameters of the model. Besides, the utilization of the pooling layer endows the sentence representation with translational invariance to features, making it more robust to positions of local features.

RNN models the sequential structure of sentence through continuously accumulating previous information of sentence[24], namely hidden states. Formally, as shown in Fig. 13, in each time step t, the hidden state $\boldsymbol{h}_t$ is dependent on the previous hidden state $\boldsymbol{h}_{t-1}$ and the present word representation $\boldsymbol{w}_t$. It can be represented as

$$\boldsymbol{h}_t = f(\boldsymbol{W}_{r1}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{r2}\boldsymbol{w}_t + \boldsymbol{b}_r) \tag{7}$$

where $\boldsymbol{W}_{r1}$ and $\boldsymbol{W}_{r2}$ are weighted matrices, and $\boldsymbol{b}_r$ is bias vector. The representation of the sentence can be defined as the final hidden state h_N, with N denoted as the length of the sentence. Several extended versions of RNN model have been proposed and applied to sentence modeling, such as Gated Recurrent Unit (GRU)[25] and Long Short-Term Memory network (LSTM)[26], with an extra gating mechanism.

Owing to the portrait of sequential structure in text, the representation learned by RNN is more sensitive to the word and phrase order in a sentence, which is crucial for semantic caption.

Transformer is a deep neural network proposed by Vaswani et al.[27], which alleviates two issues of RNN model: one is the long-distance dependence problem which means previous information will be depleted for a long sentence, another is the incapability of parallel training due to the sequential modeling. Instead of the sequential dependence, Transformer proposes a multi-head self-attention mechanism to directly connect the hidden state in each time step, as shown in Fig. 14, which can store the information of a sentence in all positions equally and be trained in parallel. Meanwhile, the multi-head mechanism can also attend to information from different vector sub-spaces. Based on this architecture, a series of pre-trained language models have been developed, with Bidirectional Encoder Representations from Transformers (BERT)[28] and Generative Pre-Training (GPT)[29] as the most representative models. They have achieved state-of-the-art performance on numerous Natural Language Processing (NLP) tasks.

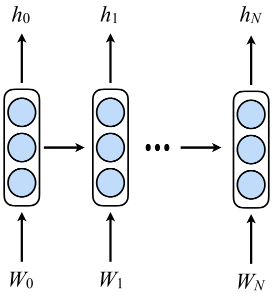


Fig. 13 Architecture of RNN.

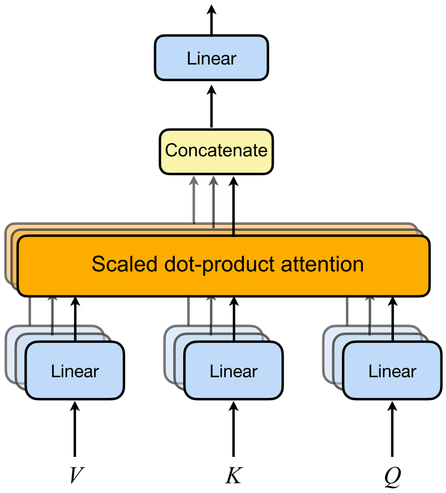


Fig. 14 Multi-head attention mechanism of Transformer. V, K, and Q denote value, key, and query in the attention mechanism, respectively.

6 Embedding-Based Network Representation

Network embedding has attracted much attention in deep learning and data mining areas since DeepWalk[30] was proposed in 2014. Before that, matrix factorization-based methods were widely adopted to project nodes in a network into real-valued vectors. In this section, we classify embedding-based network representation methods into matrix factorization-based and neural-based ones, as shown in Fig. 15.

6.1 Matrix factorization based methods

Matrix factorization based methods usually set up an optimization objective, which can be reformalized in matrix form, and then solve the optimization by eigenvector decomposition. We will introduce Laplacian Eigenmap[31] as a representative of these methods.

Given graph $G = (V, E)$, where V is the vertex set and E is the edge set, Laplacian Eigenmap[31] aimed to minimize the sum of the distances of all connected nodes, where the distance between two nodes is measured by Euclidean distance of their embeddings,

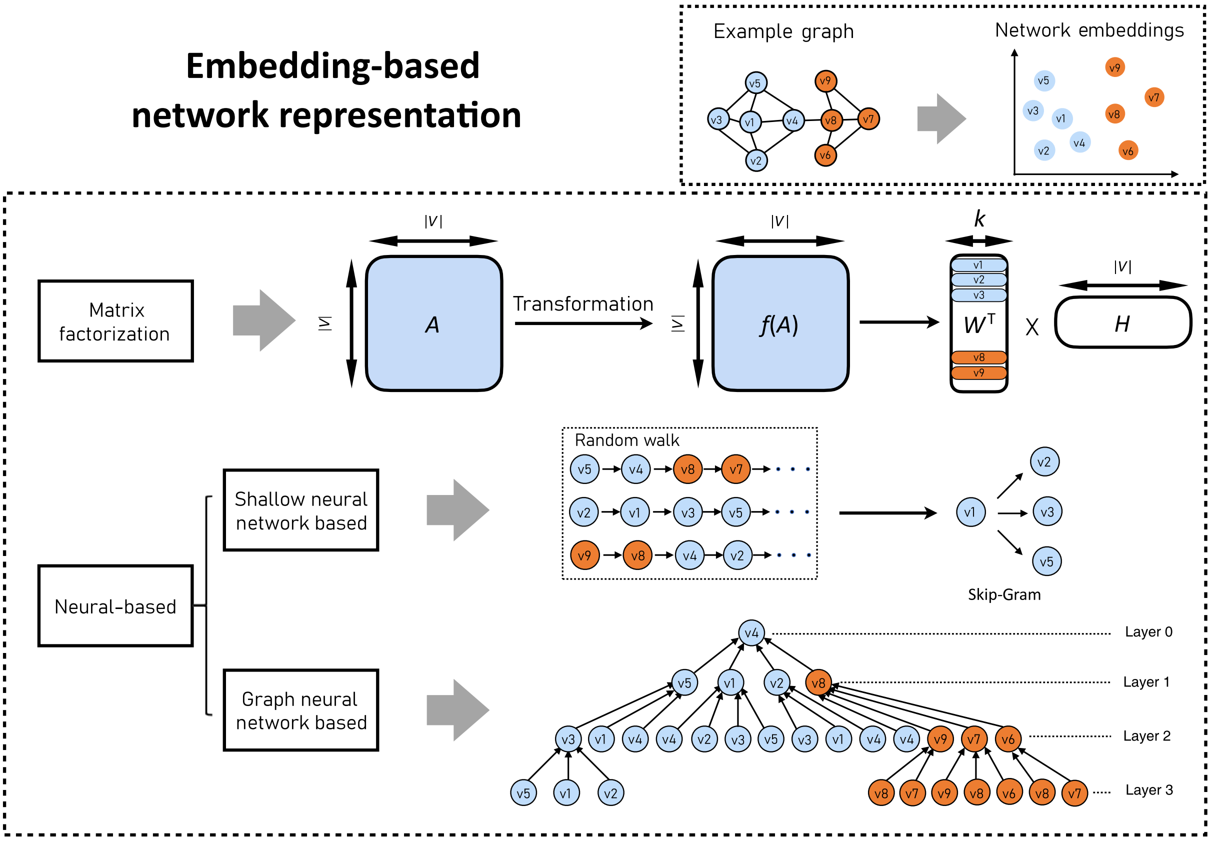


Fig. 15 An illustration of embedding-based network representation.

$$\sum_{(v_i,v_j)\in E} \|\boldsymbol{v}_i-\boldsymbol{v}_j\|^2 \tag{8}$$

where $\boldsymbol{v}_i$ is the embedding of vertex v_i.

Assume that $\boldsymbol{R}$ is a $|V|$-by-d matrix, where the i-th row of $\boldsymbol{R}$ is the d-dimensional embedding $\boldsymbol{v}_i$ of node v_i. Laplacian Eigenmap added a constraint to avoid the trivial all-zero solution,

$$\boldsymbol{R}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{R}=\boldsymbol{I}_d \tag{9}$$

where $\boldsymbol{D}$ is the $|V|$-by-$|V|$ degree matrix, D_{ii} is the degree of node v_i, and $\boldsymbol{I}_d$ is the d-by-d identity matrix. Then the optimal solution of $\boldsymbol{R}$ is proved to be the eigenvectors with d smallest nonzero eigenvalues of Laplacian matrix $\boldsymbol{L}$, i.e., the difference of diagonal matrix D and adjacency matrix $\boldsymbol{A}$.

During the last decade, gradient descent techniques are also used to solve the optimization problem in matrix factorization instead of eigenvector decomposition, especially when the close-form solution does not exist. Gradient descent techniques make it easier to train matrix factorization-based methods and help this line of work get popular. By the way, the topic model introduced in sentence embedding methods can also be viewed as a general factorization process of the document-word cooccurrence matrix.

6.2 Neural-based methods

Neural-based methods can take advantage of neural networks as well as deep learning techniques to build their optimization objectives. Their model could be deep or non-linear, and thus more flexible than the matrix factorization-based ones. Therefore, neural-based methods have become the mainstream for learning network embeddings in recent years. We further categorize relevant methods into shallow neural network based and graph neural network based ones.

6.2.1 Shallow neural network based methods

Now we will first introduce three popular unsupervised network embedding algorithms, i.e., DeepWalk, node2v ec, and LINE. Then we will briefly illustrate the idea of graph neural networks, a powerful neural architecture to encode structural information and feasible for supervised or semi-supervised end-to-end training.

DeepWalk. Inspired by the great success of word2vec[15], as shown in Table 1, DeepWalk[30] makes an analogy between word/sentence and node/random

Table 1 Analogy between word2vec and DeepWalk.

Method	Object	Input	Output
Word2Vec	Word	Sentence	Word embedding
DeepWalk	Node	Random walk	Node embedding

walk, and adopts word2vec algorithm for learning node embeddings. The intuition behind is that node frequency in short random walks and word frequency in documents both follow power law.

Formally, a random walk $(v_1, v_2, \ldots, v_i)$ is a node sequence started from node v_1, and each node v_k is randomly selected from the neighbors of node v_{k-1}. Random walks have been used in many network analysis tasks, such as similarity measurement[32] and community detection[33]. Therefore, the structural information can be encoded into sampled random walks.

Then DeepWalk treats sampled random walks as sentences from a text corpus, and employs Skip-Gram and hierarchical softmax model for learning node embeddings. The overall objective function can be obtained by summing up every node in every sampled random walk.

By preserving structural information in learned node embeddings, DeepWalk outperforms traditional symbol-based representations, such as adjacency matrix, on both efficiency and effectiveness by alleviating the computation and sparsity issues. Besides, compared with the adjacency matrix, random walks can better characterize the network structure by capturing the similarity between the nodes that are not directly connected. Thus we can achieve better performance on downstream tasks with more structural information provided.

Node2vec. Note that DeepWalk generates random walks by choosing the next node from a uniform distribution. Node2vec[34] further generalizes DeepWalk with Breadth-First Search (BFS) and Depth-First Search (DFS) on random walks. Specifically, node2 vec proposes a neighborhood sampling strategy for generating random walks and can smoothly interpolate between BFS (microscopic local neighborhoods) and DFS (macroscopic community information).

Formally, given a random walk arriving at node v through edge (t, v), node2vec defines the unnormalized transition probability of edge (v, x) for next walk step as $\pi_{vx} = \alpha_{pq}(t, x)$, where

$$\alpha_{pq}(t,x) = \begin{cases} \frac{1}{p}, & \text{if } d_{tx} = 0; \\ 1, & \text{if } d_{tx} = 1; \\ \frac{1}{q}, & \text{if } d_{tx} = 2 \end{cases} \tag{10}$$

and d_{tx} denotes the shortest path distance between node t and x. p and q are controlling hyper-parameters: a small p will increase the probability of revisiting and restrict the random walk in a local neighborhood, while a small q will encourage the random walk to move to distant nodes. The operations of node2vec after the generation of random walks are the same as DeepWalk.

LINE. LINE[35] parameterizes first-order and second-order proximities between vertices for learning network embeddings. The first-order proximity denotes nodes, that are directly connected, and second-order proximity represents nodes that share common neighbors.

Formally, LINE models the first-order proximity between node v_i and v_j as the probability,

$$p_1(v_i, v_j) = \frac{1}{1 + \exp(-\boldsymbol{v}_i \cdot \boldsymbol{v}_j)} \tag{11}$$

The target probability is defined as the weighted average $\hat{p}_1(v_i, v_j) = w_{ij} \Big/ \sum_{(v_i,v_j)\in E} w_{ij}$, where w_{ij} is the edge weight. The optimization objective is to minimize the distance between parameterized probability p_1 and target probability $\hat{p}_1$,

$$\mathcal{L}_1 = D_{\mathrm{KL}}(\hat{p}_1 \,\|\, p_1) \tag{12}$$

where $D_{\mathrm{KL}}(\cdot\|\cdot)$ is the KL-divergence between two probability distributions.

For modeling the second-order proximity, the probability that node v_j appears in v_i's context (i.e., v_j is a neighbor of v_i) is parameterized as

$$p_2(v_j|v_i) = \frac{\exp(\boldsymbol{c}_j \cdot \boldsymbol{v}_i)}{\sum_{k=1}^{|V|} \exp(\boldsymbol{c}_j \cdot \boldsymbol{v}_i)} \tag{13}$$

where $\boldsymbol{c}_j$ is the context embedding of node v_j. Given two nodes sharing many common neighbors, their embeddings will have large inner products with the context embeddings of common neighbors. Therefore, their embeddings will be similar and thus can capture the second-order proximity.

Similar to Eq. (12), the target probability is defined as $\hat{p}_2(v_j|v_i) = w_{ij} \Big/ \sum_k w_{ik}$, and the optimization objective is to minimize

$$\mathcal{L}_2 = \sum_i \sum_k w_{ik} D_{\mathrm{KL}}(\hat{p}_2(\cdot, v_i) \| p_2(\cdot, v_i)) \tag{14}$$

The first-order and second-order proximity embeddings are learned independently. After the training phase, we can concatenate them as node embeddings.

6.2.2 Graph neural network based methods

Graph Neural Network (GNN) can be seen as a special kind of convolutional neural network that operates on graphs. There are three common points between GNN and CNN: local connection, shared weights, and multi-layer architectures. Each sliding window in a CNN becomes the enumeration of every node's neighborhood in a GNN, i.e., a node and all its neighbors. Therefore, in each layer of GNN, every node will update its embedding by aggregating the embeddings of its neighbors as well as itself in the previous layer. Weight matrices and non-linear functions are also employed in the update process. Taking one of the most widely used GNN architectures, Graph Convolutional Neural network (GCN)[36], as an example, the update rule in the t-th layer of GCN can be formalized as

$$\boldsymbol{H}^{(t)} = f(\boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{A} \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{H}^{(t-1)} \boldsymbol{W}^{(t)}) \tag{15}$$

where matrix $\boldsymbol{H}^{(t)}$ indicates the embeddings of all the nodes in a network, $\boldsymbol{D}$ is the degree matrix, $\boldsymbol{A}$ is the adjacency matrix with self-loops, and $\boldsymbol{W}^{(t)}$ is the trainable weight matrix in the t-th layer. The output embeddings can be directly fed into classifiers for an end-to-end training process.

7 Applications in Computational Social Science

Computational social science has received widespread attention after decades of development. As a typical inter-disciplinary area, it is involved in multifarious disciplines, including not only five primary sub-disciplines of traditional social science, namely sociology, anthropology, psychology, politics, and economics, but also other disciplines of humanities, such as linguistics, communication, and geography. Hence, we choose three of the most cited and prestigious multidisciplinary academic journals: *Nature* (We only choose the articles in the social science subject of *Nature* as candidate pool to ensure the relevance), *Science* (Articles in the main journal and the sub-journal *ScienceAdvances* are considered to ensure representativeness and relevance as well), and *PNAS* (Papers from its social science category are examined, of which the link is https://www.pnas.org/category/social-sciences), to investigate the applications of symbol-based and embedding-based representations in CSS in recent ten years (2011–2020). Specifically, we first sort published papers in these journals by the number of citations in each year (The number of citations is crawled from Bing search engine), since we believe the number of citations is an important indicator of the influence and representativeness of an article. Afterwards, top-cited papers, utilizing one or more types of symbol-based or embedding-based representations in CSS each year, are selected for our survey. Note that referring to the number of citations of papers published in recent years (i.e., 2019 and 2020) makes less sense, therefore we list all relevant papers in 2019 and 2020 for our analysis.

Since CSS is a highly intertwined discipline between social science and computer science, we further examine the number of applications using these two representations in computer science. We select 3 top conferences closely related to CSS in computer science, namely ACL (Association for Computational Linguistics), WWW (International World Wide Web Conference), and KDD (International Conference on Knowledge Discovery and Data Mining), involving the research areas of natural language processing, data mining, and network analysis. We follow the similar settings in the above three journals and choose top-cited papers each year between 2011 and 2020 for text and network.

Considering the comprehensiveness of the audience and the diversity of the topics, we highlight the representative applications in three journals in the main text. An overview of all the applications in all three journals and three top conferences is presented in the Appendix.

In this section, we first formalize the main tasks utilizing text and network data in CSS. Afterwards, we group the applications following their task formalizations, and present how existing studies utilize symbol-based and embedding-based representations to serve these tasks, in order from symbols to embeddings, and texts to networks. At last, we further summarize and compare the advantages and disadvantages between these two kinds of representations according to their applications.

7.1 Task formalization

In spite of the explosive growth in research topics, applications employing text and network data in CSS can be summarized mainly in eight prototypical tasks, i.e., description, relation, similarity, clustering, classification, regression, language model, and ranking. A simple illustration of these formalized tasks is shown in Fig. 16 . In the following, we will give explicit definitions of them, respectively.

Description denotes quantitative depiction of characteristics of data, including frequency, distribution, etc. Distinguished from inferential statistics, it is a direct summarization of the observed data, while inferential statistics aims to infer the properties of a larger population based on the analysis of observed data.

Relation aims to measure the relationship between two variables, with correlation and causality as the most typical relationships. Given certain circumstances, one variable changes, another variable also moves, then these two variables are correlated. Causality can be regarded as a kind of continuous and stable correlation, regardless of whether other variables exist and how they change.

Similarity aims to measure if two objects have similar characteristics, such as semantics, sentiment, or styles. From the technical perspective, this task is the basis for many other tasks, such as clustering and classification.

Clustering is to group a set of objects, so that objects in the same group (i.e., cluster) are more similar than objects in other groups. Note that it automatically explores the features of different categories existing in data, without requirement of the specific definition of each category.

Classification focuses on classifying each object into one or multiple specific categories in line with its properties. Different from clustering, these categories are manually defined in advance.

Regression is similar to the classification task, with the difference existing in that regression that focuses on predicting a continuous target for each object, rather than a discrete category.

Language model is a unique task for text analysis, which calculates the probability distribution over sequences of words. It is generally implemented by calculating the conditional probability of a word given its context. Taking the word sequence {'I', 'love', 'my', 'mother'} in Fig. 16 as an example, it behaves as

$$P(\text{'I'}, \text{'love'}, \text{'my'}, \text{'mother'}) = P(\text{'I'}) \times P(\text{'love'}|\text{'I'}) \times \cdots \times P(\text{'mother'}|\text{'I'}, \text{'love'}, \text{'my'}) \qquad (16)$$

where $P(\text{'mother'}|\text{'I'}, \text{'love'}, \text{'my'})$ denotes the conditional probability of predicting word 'mother', given words in the previous context subsequence {'I', 'love', 'my'}.

Ranking is a task mainly for network analysis, aiming to find out the most important or influential nodes in a network. In other words, we need to score the nodes and rank them for our purpose.

Based on the above task formalizations in CSS, we can gain an overview of the scenarios in which symbol-based

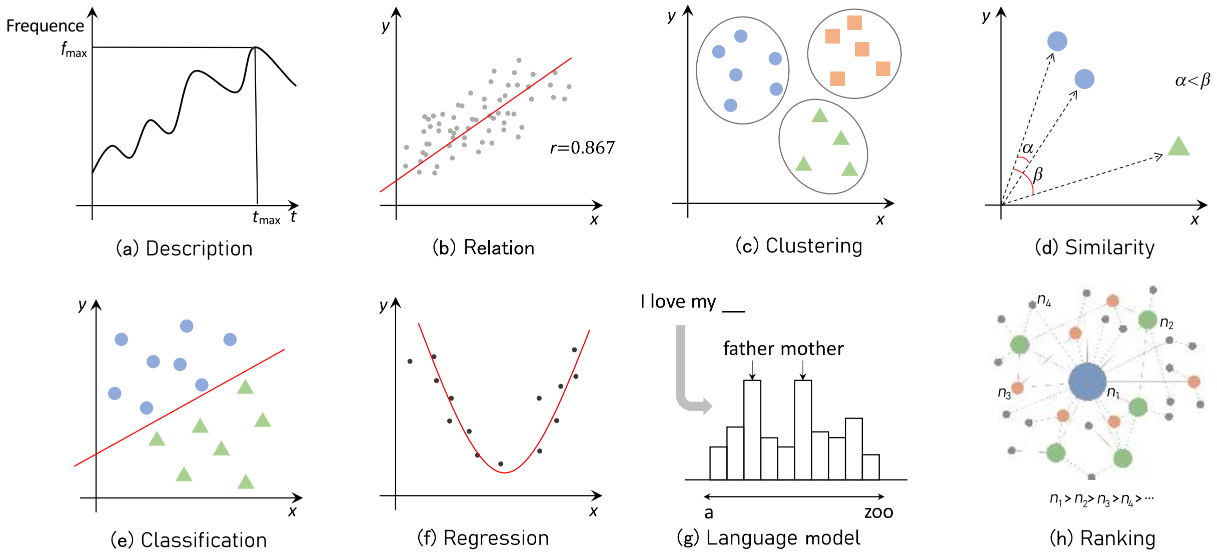


Fig. 16 A simple illustration of eight prototypical tasks.

and embedding-based representations can be applied, to further consider which type of tasks they are expert in.

7.2 Applications of symbol-based text representation

As a traditional way, symbol-based representation has been widely applied in CSS over the past decade when analyzing text data. In this section, we sort out the applications according to the type (i.e., word or sentence) of representations they employ and prototypical tasks they are formalized into. Top half of Table 2 lists the sorted applications using symbol-based text representations.

7.2.1 Applications of symbol-based word representation

Symbol-based word representation is mainly applied to the task forms of description, relation, similarity, clustering, with a few in classification and regression.

Description. Symbol-based word representation is extensively used for description, especially frequency-based word representation, owing to its intuitiveness and interpretability.

Researchers usually define specific words as representatives of an abstract concept, such as culture, linguistic grammar, and sentiment, and demonstrate the development and variations of the concept by observing their frequency changes across time and space. Michel et al.[38] tracked the words expressing time, such as “1880” and “1973” in millions of digitized books from 1800 to 2000, and found that people forget past faster as time goes by, with “1973” declined to half its peak three times faster than “1880”. Similarly, it also found that we absorb the technology faster than before with words of the invention widespread more rapidly. Yang[37] counted the frequency of two determiners “a” and “an” paired with nouns in the data where young children learn American English, respectively, and calculated the empirical probabilities of nouns co-occurring with these two determiners. Compared with expected probabilities, it discovered that young children’s language is equipped with a productive grammar rather than memorization of caregivers’ speech. Lupia et al.[40] calculated the most distinguished words co-occurring with the “National Science Foundation” or “NSF” between the Republicans and Democrats according to the count of words, and further found their different concerns for NSF.

Bruch and Newman[39], Sheshadri and Singh[41], and Golder and Macy[42] all extracted sentiment words from the corpus, and regarded the frequency of them as the indication of individuals' mood or the framing polarity of news.

Outside of frequency-based word representation, feature-based representation is also applied to the task of description, although it generally requires a large amount of manual effort. Dodds et al.[10] manually labeled happiness value of 10 000 most common words in 10 languages, and derived that a universal positive bias exists in natural language through observing the distributions of happiness scores across different languages.

Relation. Symbol-based word representation is also frequently utilized to investigate the relationship between two variables, including the correlation and causality. Alanyali et al.[43] counted the frequency of a company’s mentions in the news, and discovered a positive correlation between the frequency and its daily transaction volume.

Table 2 Applications of symbol-based and embedding-based text representations.

Representation			Task						
			Description	Relation	Similarity	Clustering	Classification	Regression	Language model
Symbol	Word	Frequency-based	[37−42]	[41, 43]					
		Feature-based	[10]	[10]				[44]	
		Network-based			[45−47]	[47, 48]	[49]		
	Sentence	Frequency-based	[50, 51]	[51]			[51−53]		[54]
		Feature-based	[55−57]	[58]	[59−62]		[45, 49, 58, 60, 63−67]	[40]	
Embedding	Word	Word embedding-based			[68−70]				
	Sentence	Topic model-based	[40, 71]	[72]	[72−74]	[75, 76]	[51, 77]		
		Neural-based			[41]		[78−80]		[79]

Dodds et al.[10] examined how the happiness scores of words vary across 10 languages and found a strong correlation between any two languages. Apart from the relationship of correlation, Sheshadri and Singh[41] investigated the causality of negative polarity of news framing and public approval, as well as legislation, where the polarity is represented by the frequency of negative sentiment words.

Similarity. Symbol-based word representation can be applied in the task of calculating the similarity between objects, where the main method is based on network-based word representation. Researchers can use network analysis methods to calculate the similarity between two networks' structures or two nodes in a network. Stella et al.[45] built two networks according to hashtag co-occurrences of two polarized groups on Twitter, and calculated the consistency of common nodes in two networks to reveal semantic similarity of two polarized groups. Different from using consistency to calculate similarity, Ramiro et al.[46] defined semantic similarity of a word's two senses based on their conceptual proximity in the network, which was constructed following the taxonomic hierarchy structure of a word form-sense dictionary. It demonstrates that a word extends its senses mainly through a nearest-neighbor chain based on the above similarity. Jackson et al.[47] constructed colexification networks of 24 emotion concepts across 2474 spoken languages, and further used Adjusted Rand Indices (ARIs) to quantify the similarity of two networks' structures. Depending on the similarity calculation, it revealed the significant difference of emotion semantics across different language families.

Clustering. Since the task of clustering is principally based on the calculation of similarity, it also commonly applies network-based word representations. We can implement clustering of words by network analysis algorithms, such as community detection. Rule et al.[48] constructed a semantic network of word co-occurrences in the annual State of the Union address (SoU) corpus, and identified discursive categories in political discourse through the community detection algorithm. Jackson et al.[47] clustered the emotion colexification networks using the community detection algorithm as well.

Classification & Regression. Symbol-based word representation has also been used in predictive tasks, mainly classification and regression. As for frequency- and feature-based representations, they usually denote as clues for objectives in a specific class or with a particular value. Huth et al.[44] represented each word in narrative stories as a vector comprising of the numbers of co-occurrences with a set of 985 common English words. Then it adopts regularized linear regression to predict the Blood-Oxygen-Level-Dependent (BOLD) responses for each subject when subjects listen to the narratives. In this manner, it reveals the semantic map across the cerebral cortex of humans. As for network-based representation, it allows for label or value propagation through the connections of nodes in the network, with the label indexing a singular category. To classify a series of hashtags into two classes: "pro-Clinton" or "pro-Trump", Bovet et al.[49] constructed a network according to hashtag co-occurrences on Twitter, with several labeled hashtags as initial seed nodes. Afterwards, it spreads labels to other hashtags in the light of connections among nodes with different labels. After several iterations, it obtains a stable label for each node in the network, namely, endows each hashtag with a fitting class.

7.2.2 Applications of symbol-based sentence representation

Symbol-based sentence representation is widely used in prototype tasks of prediction, mainly classification, while also introduced into the tasks of similarity, description, and relation. Besides, it also can be employed in the exclusive task for text, namely language model.

Classification. Symbol-based sentence representation is well received in classifying text through defined inputs. It occurs two main branches in our investigation, namely attitude classification and content classification, which we will describe below.

As for attitude classification, it covers the classifications of sentiment, emotion, and stance, etc., expressed from the text. In this branch, frequency-based and feature-based representations are often utilized jointly, while both can be used separately. Concerning frequency-based representation alone, Eichstaedt et al.[51] leveraged the unigrams and bigrams to represent posts on Facebook to predict posting users' depression status. Green et al.[52] used the same frequency-based representations to classify the partisanship of tweets' authors and further examined the polarization in elite communication about the COVID-19 pandemic. With regard to feature-based representations, Kramer et al.[63], Brady et al.[64], Jones et al.[65], and Kryvasheyeu et al.[58] all adopted dictionary-based features to study the sentiment or emotion of posts on social media, with LIWC the most well-known. Besides dictionary-based features, Stella et al.[45] incorporated lexical features, such as emoticons and acronyms, and syntactic features such as polarity shifts due to connectors to decide the sentiment of tweets. Catalini et al.[60] considered lexical features, such as part-of-speech, to represent each citation of a paper (i.e., sentences that contain the reference to another paper), and assigned the citations to two types of interest: objective and negative. When it comes to jointly use these two kinds of representations, Del Vicario et al.[66] integrated n-grams, TF-IDF, etc. (frequency-based), with emoticons, negations, and sentiment words from a predefined dictionary, etc. (feature-based), to classify the emotion of posts on Facebook. Bovet et al.[49] extracted BOW (frequency-based), hashtags, and emoticons, etc. (feature-based) to represent tweets, and investigated users' stance for Clinton and Trump in the context of 2016 US presidential election.

As for content classification, it aims to classify the text according to its substantive things, such as topic and meaning. Bakshy et al.[53] applies frequency-based representations (i.e., unigrams, bigrams, and trigrams) to classify news into "hard" (e.g., national, politic, or world affairs) or "soft" (e.g., sports, entertainment, or travel) content. Alizadeh et al.[67] combines feature-based representations with frequency-based representations to distinguish influence operations from organic activity in social media, which contains URL and words in LIWC dictionary appeared in a tweet, in addition to unigrams and bigrams.

Similarity. The similarity of two sentences or documents is usually regarded as the agreement degree of their extracted predefined features, when applying symbol-based sentence representation. Researchers normally adopt features from an existing dictionary or design features from a customized vocabulary. For example, Klingenstein et al.[59] represented each trail as the probability distribution over synonym sets in Roget' s Thesaurus, and then calculated the divergence between violent and nonviolent trials using Kullback-Leibler (KL) divergence. It shows that trials for violent and nonviolent offenses become progressively distinct through analysis of 150 year of oral testimony in the English criminal justice system. Besides, Boyd et al.[61] and Hughes et al.[62] studied the stylistic similarity of literature by representing each literary work as the distribution over a list of defined content-free words, while Boyd et al.[61] focused on the structural similarity but through the representation of distribution over LIWC dictionary. Besides above dictionary-based sentence representation, frequency-based sentence representation can also be employed, though the dimension of the representation could be relatively large. Citron and Ginsparg[50] represented each scientific article based on 7-grams occurred in the article, and investigated the text reuse in scientific corpus through calculating overlapping 7-grams between any two articles.

Description. Symbol-based sentence representation can also be used to describe the data. It generally relies on the statistics of some artificially defined features to disclose some phenomena, different from symbol-based word representation relying on frequency-based manner mostly. Jordan et al.[56] defined two scores of psychological processes: analytic thinking and clout in language of political leaders and cultural institutions, based on the statistics of function words in LIWC dictionary appeared in their text. It is derived from the fact that people's thinking and attention patterns are reflected in their use of function words. Similarly, Frank et al.[57] defined the happiness score of a tweet depending on the usage of words in the labMT dictionary. Futrell et al.[55] leveraged the syntactic features through calculating the dependency lengths of sentences across 37 languages, and unearthed that dependency length minimization is a universal property of languages.

Relation. Because of the intuitive and interpretable nature of the symbol-based sentence representation, it can be used with confidence to detect relationships between internal variables of sentences or with other external variables. For instance, Eichstaedt et al.[51] examined the use of words from LIWC in tweets, and observes the association of these features with users' depression status who post them.

Language model. To alleviate the data sparsity problem caused by the exponentially many sequences, language model generally refers to the n-gram language model in a symbol-based manner. It is assumed that the probability of the word occurred after the context history can be approximated by the probability of the word occurred after the preceding $N-1$ words, namely independent of words before these N words. Therefore, the calculation of language model depends on the frequency of n-gram occurring together in the corpus, which is widely used to measure the creativity and information presented in a sentence. For example, Piantadosi et al.[54] quantitied the information provided from a word by calculating the n-gram language model across 10 languages, and revealed that information content predicts word length better than frequency.

7.3 Applications of symbol-based network representation

Symbol-based network representation is still the mainstream used in CSS applications. The top half of Table 3 lists the applications using symbol-based network representations. We split them into the representations of node and subgraph.

7.3.1 Applications of node-based representation

Symbol-based node representations except network centrality are mainly applied to the task of description/relation, where qualitative/quantitative connections between data characteristics and a specific phenomenon or property are discussed. In contrast, centrality-based representations naturally fit the ranking task.

Description. Some work explored node-based statistics or designed indices to describe the patterns of network data. Grinberg et al.[81] studied fake news on Twitter during the 2016 US presidential election, with the help of the co-exposure network, where nodes are news websites and edges are shared-audience relationship. They employed a number of node-based simple statistics and designed indices, mostly percentage ratios, to draw their conclusions, e.g., only 1% of individuals account for 80% of fake news source exposures. Li et al.[82] employed simple node-based statistics, such as degree distributions, to describe the patterns of large mobile phone calling networks. Dankulov et al.[101] computed and visualized node-based temporal indices (e.g., the distributions of the interactivity time for users and tags), including complex probabilistic ones, to describe the dynamic patterns of users and tags in a questions & answers system.

Relation. Studying the correlation between two factors or variables is the most popular task in network analysis. Regression analysis and correlation coefficients are the most used mathematical tools for quantifying the relations.

For regression analysis, Grinberg et al.[81] studied fake news on Twitter during the 2016 US presidential election. They employed a regression model to show the relation between two variables, e.g., the sharing of content from fake news sources (as a binary variable) was positively associated with tweeting about politics. Turetsky et al.[83] studied students' peer social network and represented each student's network positions by a number of centrality-based indicators (degree, betweenness, closeness, etc.) and simple statistics. They built multiple regression models between the indicators and whether a student is perturbed by a psychological intervention. As a result, they found the intervention has positive social effects. Apicella et al.[84] characterized the social network of the Hadza hunter-gatherers in Tanzania, which may reveal the behaviours of early humans. They used regression analysis to evaluate the relationship between personal characteristics (sex, age, height, etc.) and degree (campmate ties and gift ties). For example, they found that taller people are more socially active and

Table 3 Applications of symbol-based and embedding-based network representations.

Representation			Task						
			Description	Relation	Similarity	Clustering	Classification	Regression	Ranking
Symbol	Node	Node & Edge-based statistics	[81, 82]	[81−91]	[85, 90, 92]		[93]		
		Centrality-based		[83]					[39, 94−100]
		Designed index	[101]	[91, 102−105]	[102]			[106]	[99]
		Probabilistic model	[101]				[107]	[108]	[108, 109]
	Subgraph	Motif-based statistics/ coefficients/index		[102, 110]	[102, 110]				
		Cluster-based statistics/ coefficients/index		[87]	[47]	[111]			
Embedding	Node & Subgraph	Matrix factorization			[112]	[113]	[112, 114]	[114]	[115]
		Neural-based			[116]		[116−119]	[117]	

attractive. Boardman et al.[85] discussed about how genetic factors (i.e., genotypes) can be predicted based on the genotype of his/her friends as well as the environment context. They also used regression analysis to detect the relationship between node-based factors and genotypes. Charoenwong et al.[103] employed regression analysis to understand the relation between social connections and the compliance with mobility restrictions under COVID-19 pandemic. Wu et al.[91] focused on the citation network, and employed regression analysis to reveal the relation between the team size and a number of statistics/designed indices. For instance, disruption percentile measures whether a team search more deeply into the past, which could be disruptive to science and may succeed in the future. They concluded that " large teams develop and small teams disrupt".

For correlation coefficients, Li et al.[82] computed the Spearman/Pearson correlation coefficients between two calling networks' node degree and edge weight distributions to analyze their sharing patterns. Clauset et al.[86] studied the inequality and hierarchy in faculty hiring networks of universities. They first constructed a network of institutions, where each directed edge represents a faculty member at one institution who received his/her doctorate from another. Then a prestige score for each institution is computed by node-based statistics. They showed that institutional prestige correlates well with the US News & World Report rankings, and concluded that institutional prestige leads to increased faculty production and better faculty placement. Eom and Jo[105] validated the generalized friendship paradox that your friends have on average more friends than you have in complex networks. In specific, they designed several indices as node characteristics and analyzed the degree-characteristic correlation.

For others, Wesolowski et al.[88] analyzed travel networks of people and parasites between settlements and regions based on mobile phone data. They identified the relation between human travel and parasite movement mainly by visualization and simple statistics.

Similarity. For the similarity task, node-based representations are usually used for analyzing the strength of links or how likely a link will appear between two nodes. Park et al.[92] measured the similarity of two nodes (i.e., the strength of the edge between them) as the frequency of bidirected mentions and the total bidirected call volume in seconds. They concluded that long-range edges are nearly as strong as those within a small circle of friends. Parkinson et al.[90] studied the social network of first-year graduate students, and scanned subjects' brains during the viewing of naturalistic movies. Through the statistics of the significance test, they showed that similar neural responses can help predict the friendship. Boardman et al.[85] employed descriptive statistics to represent genetic and social factors, and demonstrated the genetic homophily (persons with the same genotype tend to be friends) by significance test. Asikainen et al.[102] studied the tendency of similar people to be connected to each other by choice homophily (measured by node-based designed index) and the strength of triadic closure (measured by motif-based designed index).

Classification. To classify a node, node-based representations are usually built by integrating the information of neighbors. Garcia[93] predicted the hidden profiles (i.e., sexual orientation and relationship status) of nonusers given the profiles of disclosing users. They formalized the problem as binary classification, and simply averaged the profiles of a nonuser's friends as the node representation for prediction. Massucci et al.[107] proposed to infer the propagation paths of perturbations in a network (e.g., the spread of epidemics). They also treated the problem as binary classification and used a probabilistic model to estimate the probability of each unobserved node being perturbed given its neighbors.

Regression. Though regression analysis is widely used in CSS for detecting the correlations, only a few work targets on the regression problem. Ganin et al.[106] studied the efficiency and resilience of transportation networks, where intersections are mapped to nodes and road segments between the intersections are mapped to links. They designed node-based indices to model the commuter flows, and constructed a regression model to estimate travel delays in 20 different urban areas, with another 20 areas for calibration. Teng et al.[108] first built a probabilistic information spreading model to characterize the behaviours of nodes and estimate the collective influence of multiple spreaders, then they identified the most influential spreaders that maximize the influence.

Ranking. Centrality indicators perfectly suit the ranking task, where most work aims at finding the most important or influential nodes in a network. Pei et al.[99]

utilized various network centrality coefficients to detect the most influential information spreaders in online social networks. To study the cultural history and discover cultural centers, Schich et al.[94] constructed a directed network of cities in Europe and North America based on migration, where the endpoints of each edge in the network represent the birth and death locations of a notable individual. Then they used PageRank centrality to identify the most influential cities. Manrique et al.[95] investigated the online network of ISIS (Islamic State) members. With the help of centrality indicators, they found that although men dominate numerically, women emerge with superior network connectivity that can benefit the underlying system's robustness and survival. Hart et al.[97] built a similarity network of 200 Iroquoian village sites dating from A.D. 1350 to 1600, and concluded the importance of a specific location in population dispersal. Fraiberger et al.[98] investigated the exhibition history of half a million artists, constructing the coexhibition network that captures the movements of art between institutions. Centrality is further employed to capture institutional prestige and help understand the career trajectory of individual artists. Bruch and Newman[39] identified the most desirable users in an online dating network by PageRank centrality. Then they conducted analysis on users' strategies, e.g., both men and women pursue partners who are on average about 25% more desirable than themselves. Medo et al.[109] developed a probabilistic model to find out the discovers who are repeatedly and persistently among the first to collect the items that later become hugely popular. They also showed that traditional centrality indicators fail in this scenario.

7.3.2 Applications of subgraph-based representation

Subgraph-based representation can be further divided into motif-based and cluster-based ones. Generally, subgraph-based representation is less popular than node-based representation in terms of both paper number and task coverage.

Relation. Both cluster-based and motif-based representations are utilized in relation analysis. Trujillo and Long[87] focused on document co-citation analysis and used χ^2 test to validate whether subject communities are related to co-citation communities. In other words, χ^2 test measures the correlation between two community assignments. Kovanen et al.[110] studied the tendency of similar individuals who participate in communications. Besides similarity analysis, they also investigate how different representations correlate, e.g., edge weights and motif counts.

Similarity. Motif-based representations are still used for the similarity analysis between two nodes, while cluster-based ones are used for characterizing more high-level similarities, such as the similarity of two networks. Kovanen et al.[110] focused on the tendency of similar individuals who participate in communications by calculating the ratio score of temporal motifs (e.g., repeated call, returned call, chains, etc.). Asikainen et al.[102] studied the tendency of similar people to be connected to each other by choice homophily (measured by node-based designed index) and the strength of triadic closure (measured by motif-based designed index). To understand the universality and diversity in how humans understand and experience emotion, Jackson et al.[47] built a network of emotion concepts (e.g., "angry" and "fear") for each of 2474 spoken languages, where two concepts are connected if their meanings appear in the same word. Then they used Adjusted Rand Indices (ARIs), which measures the alignment of two cluster assignments, to quantify the similarity of two networks (i.e., languages).

Clustering. Cluster-based coefficients naturally fit the need of clustering. But most work only used the clustering of a network as their intermediate products. Therefore, they did not develop their own clustering algorithms, but directly employed traditional community detection methods instead. Thus we only present one example work here. Modularity is defined as the number of edges within given clusters minus the expected number in a network with edges placed at random, and can characterize to what extent a network can be divided into clusters. Expert et al.[111] specialized the modularity to spatial networks (e.g., road networks and location-based social networks), in order to discover space-independent communities.

Ranking. Most work studied the ranking of nodes in a network, and thus cluster-based representations are rarely used. Waniek et al.[100] studied an interesting problem: can individuals or groups actively manage their connections to evade social network analysis tools? Here each node's importance is measured by centrality, and each community's concealment is measured by a manually designed cluster-based index. They showed that simple heuristic strategies are effective to hide from the above measurements.

7.4 Applications of embedding-based text representation

With the rapid development of natural language processing and deep learning, embedding-based text representation receives increasing attention from social scientists and computational scientists. In the following subsections, we will describe the applications employing embedding-based word and sentence representations and present them according to their formalized tasks. The bottom half of the Table 2 lists the sorted applications using embedding-based text representations.

7.4.1 Applications of embedding-based word representation

Owing to the excellent performance in capturing the semantic relation, embedding-based word representation is popularly introduced into the task of similarity.

Similarity. Different from symbol-based word representation, the semantic similarity of words is reflected by the distance of word embeddings in the vector space, not based on symbol matching, so it can be used to measure the similarity of the abstract concepts. For example, Garg et al.[68] and Caliskan et al.[69] both computed the average distance between word embeddings of words denoting genders and a series of words indicating occupations, and viewed the difference between men and women as the indicator of occupational stereotypes. They compared the occupational bias reflected in the embeddings with occupation participation rates and stereotypes investigated in the traditional survey, and identified a strong association between them. Besides, it can also be utilized to expand words outside of our knowledge with similar semantics. Sivak and Smirnov[70] used word embeddings to detect the similar words of "son" and "daughter", and investigated public mentions of them on social media. It found that both men and women mention sons more frequently than daughters in their posts, which reveals that gender inequality may start early in life.

7.4.2 Applications of embedding-based sentence representation

Sentence representations obtained from topic models and neural network models are quite different in learning mechanisms and applied tasks, though both representations are based on embeddings. Therefore, we will present the applications of these two types of sentence representations separately.

For topic models, although each of its dimensions is still unintelligible, we can infer the meaning of each dimension of the representation by its probability distribution over the word list and further artificially define it as a specific topic of the text. Therefore, it has been used in various tasks of similarity, clustering, classification, description, and relation.

Similarity. Topic models represent a sentence as a distribution over a series of topics, so researchers can measure the similarity of text in semantic topics. Farrell[73] investigated the similarity of contrarian organizations' text and text from media and politics in the climate change counter-movement by LSA model, and found growth in the semantic similarity between them from 1993 to 2013. Bokányi et al.[72] also applied LSA to study the language use patterns of counties in the USA, and mined the similarity between these counties in the semantic space. To measure the linguistic distinctiveness of the context where the child produces a word, Roy et al.[74] utilized LDA model to extract the topic distribution for each first appeared word, and used KL-divergence to compare it with the background topic distribution.

Clustering. Based on the topic model based representations, we can cluster these sentences based on topics. Curme et al.[75] clustered Wikipedia into 100 different semantic topics by LDA model, and quantified the search volume of these topics in Google search engine before stock market moves. Farrell[76] used STM to obtain representations for written and verbal texts produced by individuals and organizations participating in climate change counter-movement, and clustered them into 30 topics, such as "CO2 is Good" and "Energy Production". Based on the clustering results, it revealed that corporate funding influences the written and disseminated texts of these organizations.

Classification. Topic model based representation is often operated as one of the features for the classification task, since it can supply the semantic information of text. For instance, Jaidka et al.[77] leveraged the representation learned from LDA model to predict the subjective well-being from Twitter. Besides, Eichstaedt et al.[51] also used LDA to represent posts on Facebook and predict the depression of users.

Description. Since each dimension's meaning of the topic model based representation can be inferred to some extent, it can facilitate the semantic description of the text. Lupia et al.[40] applied STM to model the statements mentioned NSF in the congressional record to find the distinctive topics of democrats and republicans, e.g., democrats care about technology and education more than republicans. Gerow et al.[71] defined the discursive influence of scholarly articles the extent to which they shape the future discourse and used the topic model to describe the influence. In other words, it estimated an article's influence as the divergence between topic distributions learned with and without this article.

Relation. Sentence representations learned from topic models have also been exploited to assess the relationship between different variables. Bokányi et al.[72] studied the relation of regional patterns in language use, socioeconomic, and cultural status of counties in the USA, such as ethnicity and tourism, where the regional pattern in language use is represented through the LSA model.

For neural network models, due to their powerful ability to fit data and capture deep semantics, neural-based representations have been gradually introduced into classification and similarity tasks in CSS. Besides, neural-based representations are also skilled at the task of language model. Below we will introduce each of them respectively.

Classification. Neural-based sentence representation is mainly applied to the classification of abstract concepts or objects of which the feature definition needs hard human efforts. As for abstract concepts, Mooijman et al.[80] used the LSTM neural network to automatically predict moral values involved in Twitter posts and suggested an association between moralization and protest violence. The complexity and ambiguity of human languages are also predicted by the LSTM neural network when investigating the languages' efficiency[79]. As for objects with hard feature definition, Fetaya et al.[78] also applied the LSTM neural network to predict the missing Babylonian text, of which the restorations require extensive expert knowledge of each genre and a large corpus of texts.

Similarity. The similarity task can be tackled by measuring the distance or similarity of neural-based sentence representations in embedding space. Sheshadri and Singh[41] utilized the Paragraph Vector Model to obtain the representations of news articles first and used cosine similarity to measure the similarity between these articles in hyperconcentrated news periods. It further demonstrated that high similarity between articles Granger causes (G-causes)[120] public attention changes and legislation.

Language model. Owing to the strength in fitting text, neural-based representations behave excellently in the language model task. LSTM neural networks are applied to construct the language model, which is a general and solid indication of language's surprisal and complexity[79].

7.5 Applications of embedding-based network representation

Lower half of Table 3 lists the applications using embedding-based network representations. Since embedding-based methods are still undergoing the emergence period in CSS, especially in the analysis of network data, only a few works on *Nature*, *S cience*, and *PNAS* adopted embedding-based representations. Hence we also add a couple of works from WWW in this subsection.

Similarity. Both of the work on similarity analysis studied the user-user friendship in a social network. Yang et al.[112] applied matrix factorization to the social network, including user-user friendship network and bipartite user-item interaction network, for learning user and item embeddings, which were employed for friend and item recommendations. Yang et al.[116] characterized a location-based social network containing both user mobility data and the corresponding social network as a hypergraph where a friendship is represented by an edge between two user nodes, and a check-in is represented by a hyperedge among four nodes (a user, an activity type, a timestamp, and a POI). Network embedding methods were then employed for both friendship and location predictions.

Clustering. Sachan et al.[113] employed topic model to build the relationship of user, community, and topic. Then they solved the optimization and computed the community distribution of each user in the social network. Note that topic models can be seen as a special kind of matrix factorization.

Classification & Regression. Embedding-based methods are widely used for the classification and regression tasks in computer science area. Besides the above mentioned methods[112, 116], Kosinski et al.[114]

applied singular value decomposition to the user-like matrix for learning user embeddings, which were further utilized for predicting private traits. Zhang et al.[117] constructed a heterogeneous network with three types of nodes, i.e., location, time, and text, from Geo-Tagged Social Media (GTSM) data. Then they jointly encoded all spatial, temporal, and textual units into the same embedding space to capture the correlations for modeling people's activities in the urban space. More recently, graph neural network based methods[118, 119] were also proposed for social recommendation, where a user friendship network and a user-item interaction network are given as input to predict future user-item interactions.

Ranking. For the ranking task, Wang et al.[115] aimed at discovering magnet communities, which are communities that attract significantly more people's interests. In detail, they learned cluster-based representations via matrix-based optimization, and ranked given communities in a domain based on their attractiveness to people among the communities of that domain.

8 From Symbols to Embeddings

Based on the introduction of applications in the previous section, we can observe that both symbol-based and embedding-based representations have been considerably adopted in CSS. To investigate their coverage definitely, we count the number of works utilizing one or both of the two representations each year, as shown in Fig. 17. By comparisons, we can find that the proportion of articles using embedding-based representations is gradually increasing over the last decade in *Nature*, *S cience*, and *PNAS*. This indicates that more and more works in CSS have considered and benefited from the embedding-based representations. We also make the same statistics in conferences of ACL, WWW, and KDD. Figure 18 shows the comparison between the numbers of applications using symbol-based and embedding-based representations in these three conferences. From Fig. 18 we can find that the number of articles using embedding-based representations has significantly exceeded those using symbol-based representations. However, compared with Fig. 17, there is a large gap between the volume of embedding-based representations in computer science conferences and the three multidisciplinary journals.

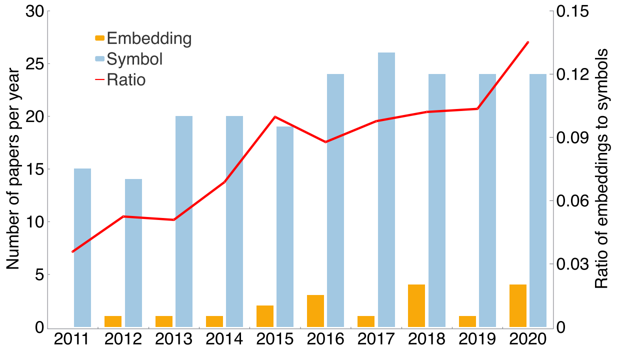


Fig. 17 Number of papers applying symbol-based representation or embedding-based representation and the ratio between them over the last decade in *Nature*, *Science*, and *PNAS*. The line is smoothed by taking a 3-year average. Detailed settings where we selected these papers are shown in Section 7.

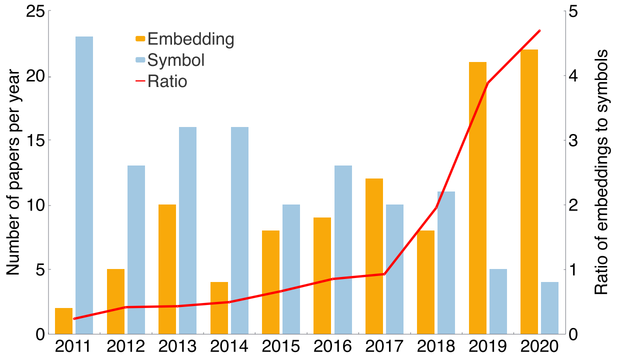


Fig. 18 Number of papers applying symbol-based representation or embedding-based representation and the ratio between them over the last decade in ACL, WWW, and KDD. The line is smoothed by taking a 3-year average. Detailed settings where we selected these papers are shown in Section 7.

This prompts us to deepen and amplify the interdisciplinary integration between social science and computer science, despite the slight shift in their research concerns.

To sum up, embedding-based representations have emerged and performed an increasingly critical role in CSS over the last decade.

We further discuss the underlying reasons for this trend and summarize the expert areas of both representations. Based on their internal mechanisms and existing applications, we conclude three key points as follows.

(1) Symbol-based representations excel at the tasks of description and relation, due to their explicitness and interpretability. Each value in the symbol-based representation denotes certain and human-readable meaning, so we can use it directly to observe the distribution of data, as well as to extract relations between objects. For example, as we introduced in Section 7.2, frequency-based word representations are applied to observe cultural changes and capture the relationship between the number of mentions in news and the stock trading volume of a company. While topic model-based representations and some neural-based representations are equipped with practical meanings to some degree[121, 122], they are still fuzzy and less compelling for researchers in social science.

(2) Embedding-based representations perform better in the tasks of prediction (e.g., classification and regression) and similarity, owing to the powerful ability of neural networks to fit the data and to extract deep semantics. On the one hand, neural networks achieve efficient input-output mapping functions through the connections of large-scale neurons. On the other hand, it realizes the extraction of deep semantics and abstract concepts by the constructions of multi-layer networks. Existing researches have demonstrated that the deep layer captures the more abstract features relative to the shallow one[123]. As presented in Section 5, abstract concepts such as social biases and moralizations are all well measured by embedding-based representations. Although we mentioned that symbol-based representations can stand for abstract concepts through some defined symbols, such representations are still partial and shallow, and hard to capture their full picture.

(3) Embedding-based representations require fewer human efforts. Symbol-based representations usually require a large amount of expert knowledge to define the features of research objects, which is labor-intensive. Besides, for some abstract concepts or objects without well-founded features, their performances will be limited. Different from them, embedding-based representations are automatically extracted from data with fewer human interventions, and even complement for human knowledge. For example, as introduced in the application section, we can use neural networks to automatically restore the missing Babylonian text, which is challenging even for experts. In addition, embedding-based representations are qualified to portray the complexity and ambiguity of the language without manual definition.

9 Discussion on Future Direction

Although the tendency from symbols to embeddings has emerged in the past ten years, there are still many challenges and open issues to be explored. Going forward, we list some essential and potential future directions involved with data representations in CSS.

(1) Pre-trained language models. In recent years, pre-trained language models have received considerable attention and achieved great success in processing textual data[28, 124]. The models learn rich semantic information from massive textual data, such as encyclopedias and books, with merely fine-tuned in downstream tasks to obtain efficient embedding-based representations. Therefore, for CSS, we can obtain more generalized and robust textual representations with the aid of pre-trained language models. The representations can not only be used to analyze social phenomena from text more extensively and accurately, but also reduce the manual annotations for those tasks requiring enormous labeled data, compared to representations learned from traditional neural network models.

(2) Graph neural networks. Through the message passing mechanism, graph neural networks[125] can effectively model both the network topology and node/edge features (e.g., text information) simultaneously, thus providing a unified framework to take advantage of information from heterogeneous sources. Many scenarios in CSS need to deal with a social network as well as individual characteristics. Therefore, graph neural network techniques have great application potentialities for CSS studies, which can learn representations integrating the information of both text and network. In fact, various applications in computer science, such as natural language processing[126] and recommendation systems[127], have already adopted graph neural networks for modeling.

(3) Design as prediction and similarity. Embedding-based representations are well-known for rich and deep semantics, while symbol-based representations are usually preserved in partial and shallow semantics. Meanwhile, embedding-based representations are skilled at the task of prediction and similarity. Therefore, to take full advantage of the strong semantics in embeddings, researchers in CSS are encouraged to design the research problem as a prediction or similarity task whenever possible. For example, we can design the problem of social bias as a similarity measurement between the embeddings of gender words and neutral words[68, 69]. In addition, the complexity of human language can be designed as a predictive task, which views the predicted probability of a word or sentence using language model as the indicator[79].

(4) Interpretability. Admittedly, a drawback of embedding-based methods is the lack of interpretability. This problem would harm the application for decision-critical systems related to ethics, safety, or privacy. Though the interpretability of embedding models, especially neural network models, has not been fully addressed yet, researchers in the computer science area have made some efforts towards better explainability of neural-based models[128]. Therefore, taking advantage of both embedding-based models and explainability analysis methods for effective and (partially) explainable predictions would be an intriguing direction.

10 Conclusion

As an emerging and promising inter-disciplinary field, computational social science has attracted considerable research interests over recent years. Two main types of data, namely text and network data, are widely used in studies of CSS. In this survey, we first summarize the data representation into symbol-based and embedding-based representations and further introduce typical methods when constructing these representations. Afterwards, we conduct a comprehensive review on the applications of these two classes of representations based on more than 400 top-cited literature from 6 classic journals and conferences. According to the statistics of these applications, a tendency that embedding-based representations of text and network in CSS are emerging and growing is discovered, which we further discuss the reason contributed to. Finally, we suggest four challenges and open issues in CSS, which are essential and potential directions to be explored.

Appendix

With the explosive growth in research topics in CSS, we divide the topics of applications we investigated into 9 domains, which is inspired from the 5+ primary sub-disciplines in traditional social science, namely sociology, anthropology, psychology, politics, economics, and other fields of humanities, including linguistics, communication, geography, and environment. Note that each work can exist in multiple domains if they are relevant simultaneously. We also list all relevant papers on the GitHub link: https://github.com/thunlp/CSSReview.

In the following, we will divide these applications into different domains, and introduce them from text to data according to the data type used, and further present them from symbol-based representation to embedding-based representation according to the representation type used.

A.1 Text

A.1.1 Symbol-based representation

Symbol-based representation of text is mostly used in the fields of sociology, followed by linguistics, psychology, geography, politics, and communication, with few utilized in economics, environment, and anthropology.

In the domain of **sociology**, a series of works utilize the symbol-based representation of text to analyze and detect misinformation and misbehaviour in online world, such as rumor[129−133], fake news[134] or image[135], low quality Wikipedia[136], hate speech[137, 138], abusive language and behaviour[139, 140], social bots[45, 141], sockpuppets[142], cybercriminal activity[143], influence operations[67], and text reuse in scientific papers[50], where they usually manipulate n-gram, BOW, TF-IDF features, as well as linguistics features, such as length, URL, hashtag in the text, accompanied with syntactic features, such as part-of-speech tagging and dependency relations. Extra lexicon, such as LIWC, is also widely used to extract keywords in the above analysis and detection works. In addition, the privacy issue is a hot topic where researchers use the symbol-based text features to prevent privacy disclosure across multiple online sites[144, 145]. These text features also benefit the search for informative posts[146, 147] and assessment of damage[148] in a disaster, while they can also contribute to the social power relation prediction[149]. Different from the above studies, a list of works aims to explore the behaviour law of human in online social media based on the symbol representations of words, for instance, hashtag adoption[150] and collective attention on Twitter[151], pursuit in online dating markets[39], and user feedback in application store[152].

In the domain of **linguistics**, a set of works focus on the study of linguistic phenomenon and trend. Some researchers count the frequencies of linguistic features, such as n-gram and emoticon, to investigate linguistic phenomenon including the evolution of grammar[38], and the correlation with socio-economic variables[153]. Taking Michel et al.[38] as an example, they counted the regular forms (added "-ed") and irregular forms (conjugated extraordinarily) of verbs from 1800 to 2000, such as "strived" and "strove" of "strive". Through the quantitative analysis, they found the linguistic fact that irregulars generally yield to regulars, with 16% of irregulars changed into regularity of more than 10%. Besides the linguistic trend, n-gram language model is used to approximate the information content of each word[54] or distinctiveness of language[154, 155].

Another set of works concentrate on the text analysis of various genres, such as debate and narrative. Boyd et al.[61] counted the function and cognitive words across each text with LIWC, to analyze the structures of narratives in different types. Jordan et al.[56] also used the LIWC to measure analytic thinking and clout in leaders' debates and speeches, and found a general decline in analytic thinking and a rise in confidence. In addition, designed linguistic lexicons, accompanied with semantic and syntactic features, are also popularly adopted in language quality detection, such as the detection of politeness[154], popularity[156], and biased statements[157].

In the domain of **psychology**, dictionary-driven text representations are widely utilized, with LIWC and LabMT[10] as mostly popular dictionaries. Kramer et al.[63] used LIWC to define the emotion of posts and found the emotional contagion through social networks. Frank et al.[57] employed the LabMT to measure the happiness expressed in language and discovered that happiness increases with distance from people's average location. Moral Foundation Dictionary is also incorporated to assist the prediction of moral values involved in Twitter posts[80]. Besides emotion and happiness, dictionary-driven representations are also extensively used to detect depression in social media[105, 158]. Despite the wide adoption of dictionary-driven representations, Jaidka et al.[77] made a comparison between unsupervised dictionary-driven and supervised data-driven methods, and verified that the latter is more robust for well-being estimation from social media data. Therefore, outside of the dictionary-driven representations, linguistic features, such as n-gram and BOW, are also applied to represent text in psychology, combined with the supervised machine learning method. For instance, Chang et al.[159] used them to distinguish a person's intention and others' perception of the same utterance, while Kern et al.[160] took them as signals for personality prediction.

In the domain of **geography**, a set of studies focus on geo-location inference, in which case the location where a textual message is generated is discovered[161–163], and route navigation, with the aim to provide a more promising route according to sentiments detected from geo-tagged documents in social media[164]. As for geo-location inference, Ikawa et al.[161] proposed a method to learn associations between a location and its pertinent keywords extracted from historical messages, while Ryoo and Moon[162] extracted the spatial correlation between texts and GPS locations from tweets with GPS-tags. Besides, Wing and Baldridge[163] adopted simple supervised approaches on the textual content of documents as well as a geodesic grid, so as to acquire the discrete representation of the earth's surface. With regard to route navigation, Wing and Baldridge[163] presented a system to recommend routes based on sentiments exposed from Twitter tweets towards places, by combining eight existing sentiment analysis tools, including LIWC, Happiness Index, SentiWordNet, SASA, PANAS-t, Emoticons, SenticNet, and SentiStrength.

In the domain of **politics** , most studies focus on investigating political activities and analyzing ideology applying symbol-based text representations. As regards political activities, Alizadeh et al.[67] utilized series of defined features, such as n-gram, URL, and LIWC, to predict social media influence operations, while Lupia et al.[40] extracted the most distinguished words and sentiment words from statements in the congressional record, to explore the congressional concern about National Science Foundation. Besides, Jordan et al.[56] used LIWC lexicon to analyze the style of political leader's language and further discussed the long-evolving political trends. With regarding to ideology analysis, Preoţiuc-Pietro et al.[165] and Bovet et al.[49] used similar symbol-based features, such as n-gram, URL, and emoticon, to predict political ideology and opinion toward presidential candidates of Twitter users, while Burfoot et al.[166] aimed to predict sentiments in congressional floor-debate transcripts with unigram features.

In the domain of **communication**, symbol-based text representation is employed to mining the content in communication. Jenders et al.[167] took hashtags, mentions, and sentiments as symbol features to predict viral tweets. Sheshadri and Singh[41] utilized *n*-gram features to analyze the news framing and explore its public and legislative impact, while Green et al.[52] adopted similar features to represent tweets sent by political elites and further analyzed the polarization in elite communication on the COVID-19 pandemic. Arous et al.[168] calculated TF-IDF scores of words as tweet features to help the detection of social influencers in communication.

In the domain of **economics**, researchers are greatly interested in revealing economical phenomenon based on the relationship between financial news and the stock market, using text-based correlational analyses[43] and the combination of several basic linguistic features[169]. Specifically, Alanyali et al.[43] adopted correlational analyses, according to daily number of mentions in the Financial Times for each company of interest, for the purpose of quantifying the relationship between decisions made in stock market and situation in financial news. Xie et al.[169] utilized scores for words in the Dictionary of Affect in Language (DAL)[170] via part-of-speech, along with bag-of-words in order to predict change in stock price according to financial news.

In the domain of **environment**, the main research interest lies in analyzing social media text data generated before, during, and after the occurrences of natural disasters, such as earthquake, hurricane, etc. In the research of Kryvasheyeu et al.[58], LIWC[9] and SentiStrength[171] were adopted for analyzing the sentiments embedded in social media texts, posted before, during, and after Hurricane Sandy, in order to investigate if the sentiment signal indicated the damage inflicted by the hurricane. Besides, Ghosh and Desarkar[172] proposed modified TF-IDF based approaches to better classify disaster related social media tweets, so that the rescue and relief operations can be better launched when natural disasters occur.

In the domain of **anthropology**, studies related to cultural evolution serve as the major interests of researchers. Specifically, two kinds of cultural shifts are studied, namely the cultural changes accompanying the monopolization of violence by the state[59] and the cultural universality and diversity in music[173]. In particular, Klingenstein et al.[59] applied a bag-of-words model as a symbol-based representation of texts to coarsely categorize the words that occur in jury trials into several predefined classes, and further analyzed the extent to which the patterns of talking in a criminal trail vary from violent to nonviolent offenses, and how these differences evolve over time. Mehr et al.[173] conducted a systematic analysis regarding the features of worldwide vocal music, where four kinds of representations were derived for each song. Using machine classifiers, they managed to observe the universality and variability in musical behaviour, reflecting cultural evolution in forms of music.

A.1.2 Embedding-based representation

Embedding-based representation of text mostly benefits the sociology, then geography, politics, psychology, environment, economics, and linguistics successively, with few adopted in communication and anthropology.

In the domain of **sociology**, embedding-based text representation is mostly adopted in content mining, misinformation, and misbehavior detection, as well as human trait prediction. As for content mining, the topic model is widely used. Singer et al.[174] adopted it to extract the topic in Wikipedia and was eager to understand why we read Wikipedia. Fu et al.[152] and Sachan et al.[113] used it to analyze content of users' writing, and discovered users' preferences and interests, while Weerasinghe et al.[175] used it to mine underlying topics of comments by pods, aiming to increase the popularity of users' content effectively. Further, Gerow et al.[71] built a dynamic topic model to measure how the content shapes future scholarship, namely its discursive influence of a paper across scholarship. Zhang et al.[117] and Wang et al.[176] incorporated extra data outside the text, such as region and time, to uncover the spatial and temporal topics. As for misinformation and misbehavior detection, word embedding methods and deep neural networks are widely used in this direction. Word embedding methods, such as Skip-Gram and GLOVE, are commonly used in social bias detection[68−70], e.g., gender bias and ethnic bias. For example, Sivak and Smirnov[70] computed the average distance between word embeddings of last names in various groups and a series of adjectives, and viewed the difference in distance between the common group and Asian group as the score for Asian bias. Besides bias detection, rumor and fake news detection are also hot topics employing embedding-based representation. They usually use RNN models as basic frameworks to encode the text representation[134, 177], with VAE[178], GAN[179], and Bayesian model[180] further improving the performance. Detection of other misinformation and misbehavior, such as toxicity triggers[181], abuse language[139], and hate speech detection[137], apply the deep neural networks to obtain the text representation as well. As for the human trait prediction, researchers endeavor to use LSTM to predict human age and gender[182], as well as activity[183], while [184] learning the representation of bios of each user with GCN to predict the user's occupation.

In the domain of **geography** , embedding representation of texts was applied in multiple application scenarios using geo-tagged social media data, including geo-location estimation[185, 186], geographical topical analysis[187, 188], urban dynamics discovery[117], human mobility modelling[189], and local event detection[190]. Specifically, topic models were employed for obtaining location-specific topics for tweets[185] and discovering language characteristics along with common topics exposed in geo-tagged Twitter streams[188]. In addition, multi-modal signals, in the form of spatial, temporal and texts, were utilized for different research purposes. For instance, Zhang et al.[117] proposed a novel cross-modal representation learning method to embed all spatial, temporal, and textual units into the same vector space in order to uncover urban dynamics. Yuan et al.[187] discovered spatio-temporal topics for Twitter users by using a probabilistic generative model for user behavior modelling from the geographic and temporal perspectives. Zhang et al.[190] presented a method to leverage multi-modal embeddings for the purpose of accurately detecting local events. Furthermore, Miura et al.[186] adopted a complex neural network, which was able to unify the representations learned from text, metadata, and user network, and an attention mechanism to better infer geo-locations of tweets.

In the domain of **politics**, existing researches can be divided into three classes using embedding-based text representations: political ideology detection, political relation extraction, and political technique analysis. Regarding political ideology detection, topic model, word embedding, and neural networks all have been adopted. For instance, Farrell[76] used STM to discover ideological polarization around climate change, and further examined the influence of corporate funding on it. Preoţiuc-Pietro et al.[165] used Word2vec to assist political ideology prediction. Hierarchical LSTM and FastText are also applied to detect political perspective[191] and stance[192]. As to political relation extraction, the topic model is mainly used to look at the relationship between republican legislators[193] or extract events between political actors from news corpora[194]. For political technique analysis, topic models and neural networks were also used to catch a glimpse of processes of framing[195], propaganda techniques[196], and political ads[197].

In the domain of **psychology**, most studies concentrate on mental health identification with embedding-based text representations. For instance, the topic model was used to mine topics from statuses and predict depression of patients[51]. RNN and CNN models were employed to represent typing data when using mobile phone and users' posts in Reddit, for mood detection[198] and suicide risk assessment[199], respectively. Besides the mental health, embedding-based text representations are also used in other psychological sphere, such as moralization[80], intention[200], and happiness[57].

In the domain of **environment** , issues related to climate change and air quality prediction attract the most attention from scholars. In Ref. [73], LSA[201] was adopted to examine the impact of different climate-contrarian organizations' ideas, regarding climate change counter-movement, on news media and bureaucratic politics. With respect to air quality, Jiang et al.[202] deployed a deep learning model, based on convolutional neural network and overtweet-pooling, on social media data to enhance air quality prediction.

In the domain of **economics**, the majority of works utilized embedding-based representations of text to investigate issues related to stock market[75, 169, 193, 203, 204], while others concentrated on electronic commerce (e-commerce)[205] and socio-economic indicators[206]. In particular, topic modelling techniques, such as LDA, were employed in a set of studies, where Liu et al.[205] managed to predict loyal buyers for e-commerce, and Curme et al.[75] quantified the semantics of search behavior of Internet users and identified topics of interest before stock market moves. In addition, scholars are also interested in combining textual contents and other types of signals for stock market related research. For instance, Nguyen and Shirai[

193] incorporated sentiment signals from social media into topic models to better predict stock price movement; Yang et al.[203] proposed a novel model architecture based on Transformer[27] to harness the textual and audio information for predicting future stock price volatility; Xu and Cohen[204] designed a deep generative model for stock movement prediction by jointly exploiting text and price signals.

In the domain of **linguistics** , embedding-based representation is mostly used in two fashions as graphical model (especially topic model) and word embedding. As for graphical model, it is utilized to capture the latent information behind the text, such as linguistic topics. Hong et al.[188] used topic model to discover geographical patterns in language use, while Bokányi et al.[72] further related the patterns to demographics. Roy et al.[74] also adopted it to capture the topic distributions of context, where children accumulate interactions and learn words. Doyle et al.[207] proposed a graphical model to model the linguistic alignment in Twitter interactions, which is an important measure of accommodation. As for word embedding, it is usually used to detect the linguistic change across corpora and time, because of the ability to capture semantics. Kulkarni et al.[208] proposed an approach to detect the linguistic change in the meaning and usage of words by Skip-Gram, while Gonen et al.[209] designed a more simple, interpretable, and stable method with Skip-Gram as well.

In the domain of **communication**, topic model is most observed to obtain text representation. Tsur et al.[195] applied it to analyze the statements from congress, which attempts to gain insights about agenda setting. Both Tang et al.[210] and Farrell[73] aimed to find topical aspects of actors and identify the most influential actors in a network. Besides, Ref. [41] employed paragraph vector to estimate similarity between two news documents, devoted to the impact of news framing.

In the domain of **anthropology** , Fetaya et al.[78] concentrated on the reconstruction of a lost ancient heritage, with the help of embedding-based representation of text. In particular, they employed recurrent neural networks to reconstruct the damaged and missing ancient Akkadian texts from Achaemenid period Babylonia.

A.2 Network

A.2.1 Symbol-based representation

In the domain of **sociology**, most studies are conducted on the social network. For online social networks, the research data usually came from popular websites or communication applications, such as Twitter[211−213], Facebook[214−216], Yahoo[217], and Wechat[218]. For offline social or friendship networks, the studied scenarios are quite diverse, such as the dating network[219], the social network structure of potential male raiders[220], problem-solving networks[221] where people worked by groups and collaborated with each other, and even the social network of cooperative bird species[222].

In terms of research problem and methodology, we summarize the following four patterns of these literature:

The first category is to study whether a phenomenon exists in the network. For example, Ref. [219] studied cross-racial communication to detect the existence of racial prejudice. Reference [223] aimed to find the social hierarchy and stratification among humans in social networks. This line of work usually employs simple statistics or proposed indices involving related features or factors for their methods.

The second category is to find out the structural patterns leading to a specific property. For example, Glowacki et al.[220] tried to find out how the formation of social network structure will lead to a potential male raider. This line of work usually analyzes the patterns of subgraphs (e.g., the frequency of specific subgraphs[221]) for modeling the correlations.

The third category is to identify the most important nodes in a network. For example, Teng et al.[108] aimed at identifying the most influential spreaders that maximize information flow. This line of work usually employs various network centrality coefficients (e.g., the degree of a node) as the measurements.

The fourth category is to predict the future behaviours of users in a network. Common scenarios include recommendation system[224] and information diffusion (e.g., the spread of rumors or misinformation[225]). This line of work needs to model the temporal dynamics and user preferences for predicting future behaviours. The detailed models are quite personalized and differ from each other.

In the domain of **anthropology**, most works employ similar symbol-based network representations for analysis, such as PageRank score or betweenness coefficient for extracting the most important nodes. Therefore, it would be more interesting to see what kind of networks they built to solve their problems.

The first kind is location networks. To study cultural history and discover cultural centers, Schich et al.[94] constructed a directed network of cities in Europe and North America based on migration, where the endpoints of each edge in the network represent the birth and death locations of a notable individual. Hart et al.[97] built a similarity network of 200 Iroquoian village sites dating from A.D. 1350 to 1600, and concluded the importance of a specific location in population dispersal. Lulewicz[226] also constructed a network of sites from the southern Appalachian region between A.D. 800 and 1650, to study the variation of Mississippian sociopolitics.

The second kind is social or friendship network. Fowler et al.[227] studied the correlation between genotypes and friendship networks, and identified a positively correlated (homophily) one and a negatively correlated (heterophily) one from all six available genotypes. Boardman et al.[85] also discussed genotypes and friendship networks, but with more consideration of environment context. Apicella et al.[84] characterized the social network of the Hadza hunter-gatherers in Tanzania, which may reveal the behaviours of early humans. A. I. Roberts and S. G. B. Roberts[228] used the social bonds between wild chimpanzees to inspire the study of human evolution.

There are also other kinds of networks. For instance, Hilger et al.[229] constructed a brain network to understand human intelligence, where nodes correspond to regions in a grey matter and edges represent high positive correlations of signals between nodes.

In the domain of **linguistics**, only a few work utilize network structure for their study. They study the networks of concepts, words or languages, and usually use simple statistics or cluster coefficients for analysis.

To understand the universality and diversity in how humans understand and experience emotion, Jackson et al.[47] built a network of emotion concepts (e.g., “angry” and “fear”) for each of 2474 spoken languages, where two concepts are connected if their meanings appear in the same word. Youn et al.[230] explored a more general problem, i.e., the universal structure of human lexical semantics. To be more specific, Youn et al.[230] built a weighted network of concepts using cross-linguistic dictionaries: sometimes a single “polysemous” word from one language can express multiple concepts that another language represents using distinct words. The frequency of such polysemies between two concepts can be seen as a measure of their semantic similarity. Sizemore et al.[231] focused on the sparsity (i.e., knowledge gaps) of semantic feature networks of humans, where words correspond to nodes and are connected by shared features, to understand the process of language learning. Ronen et al.[232] constructed a co-spoken language network to figure out the influence of different languages.

In the domain of **psychology**, the most widely used network types are the brain and social networks.

For brain networks, Taruffi et al.[233] found that compared with happy music, sad music is linked to greater centrality of the nodes of the default mode network (i.e., a set of brain regions typically active during rest periods). Schmälzle et al.[234] studied the functional connectivity of brain regions under social inclusion or exclusion.

For social networks, Turetsky et al.[83] showed that psychological interventions can strengthen the connections of peer social network in terms of node degree, closeness, betweenness, etc. Morelli et al.[235] studied how psychological traits correlate to centrality in social networks, and concluded that people high in well-being are central to the “fun” networks, while people high in empathy are central to the “trust” networks. Kim et al.[236] showed that occupying a bridging position in a social network may alleviate the impact of depressive symptoms among older men, whereas the opposite holds true for older women. Ito[237] studied how networks of general trust will affect the willingness to communicate in English for Japanese people, via the analysis of centrality indices.

In the domain of **geography**, the wide adoption of mobile devices significantly benefits the collections of location-based data, and thus facilitates relevant researches in this area. There are two main types of networks discussed in this work: transportation network and location-based social network.

Transportation networks include road/street networks and travel/mobility networks. For example, Bao et al.[238] utilized bike trajectory data on road networks to develop bike lane construction plans. Ganin et al.[106] studied the efficiency and resilience of transportation networks, where intersections are mapped to nodes and road segments between the intersections are mapped to links. Barrington-Leigh and Millard-Ball[239] analyzed a time series of street network and discussed road building in new and expanding cities for urban development. Taking the London rail network as an example, Yadav et al.[240] found that topological attributes designed for maximizing efficiency in urban transport networks will make the network more vulnerable under intense flood disasters. On the other hand, Wesolowski et al.[88] analyzed travel networks of people and parasites between settlements and regions based on mobile phone data. Bonaccorsi et al.[241] studied the effect of lockdown restrictions on the economic conditions of individuals and local governments based on the Italian mobility network. Santi et al.[242], Vazifeh et al.[243], and Liu et al.[244] focused on vehicle-shareability networks for better taxi or bike-sharing services. Riascos and Mateos[245] also analyzed taxi trip data and built a directed weighted origin-destination network for the study of long-range mobility. There are also some work[246, 247] proposed for data collections of transportation networks.

A location-based social network can be either online or offline. For online networks, Ref. [248] found that two individuals' movements strongly correlate with their proximity in the social network. Cho et al.[249] tried to understand the basic laws of human motion and dynamics based on location-based social networks and cell phone location data. Li et al.[250] focused on profiling users' home locations in the context of a social network. Yang et al.[251] proposed Socio-Spatial Group Query (SSGQ) to select nearby attendees with close social relation based on users' social networks on Facebook, as well as their spatial locations from Facebook check-in records. For offline networks, Sun et al.[252] studied a time-resolved in-vehicle social encounter network on public buses in a city. Sekara et al.[253] explored the dynamic social network of about 1000 individuals and their interactions measured via Bluetooth, telecommunication networks, or online social media, etc.

In addition, other relevant works study infrastructure networks of urban microgrids[254] or the community detection problem on general spatially-embedded networks[111].

In the domain of **economics** , the most discussed networks are financial institution networks and trade networks.

For financial institution networks, Battiston et al.[255] studied the multi-layer networks of financial institutions connected by contracts and common assets, and showed that the complexity of financial networks may increase the social cost of financial crises. Bardoscia et al.[256] also studied the network of financial institutions, and discussed how the instability of model ecosystems is relevant to the dynamical processes on complex networks.

For trade networks, Porfirio et al.[257] studied the structural changes in the global agricultural trade network under greenhouse gas emissions. It is worth noting that Porfirio et al.[257] employed matrix factorization, a technique widely used in network embedding learning, for their modeling. However, they only utilized the singular values and discarded the vectors in singular value decomposition. Thus, we still classify this work as a symbol-based one. Ren et al.[258] characterized the international trading system with a multi-layer network with each layer representing the transnational trading relations of a product. They studied a nation's economic growth by analyzing node degrees and product rankings over time.

For others, Anderson[259] built a network of skills based on their relationship in the market, and showed that workers with diverse skills can earn higher wages than those with more specialized skills. Bonaccorsi et al.[241] studied the effect of lockdown restrictions on economic conditions of individuals and local governments based on the Italian mobility network.

In the domain of **politics** , most researches are conducted on social media or online social networks, and discuss ideology or elections.

For the ideology topic, Farrell[76] constructed an organization network and concluded that organizations with corporate funding are more likely to write and spread texts that lead to ideological polarization on the climate change issue. By analyzing the retweeting behaviours in online social networks, Brady et al.[64] found that the expression of moral emotion is key for the spread of moral and political ideas or ideology. Bail et al.[260] extracted a following network of 4176 opinion leaders on Twitter, and used the first component of the adjacency matrix to create liberal/conservative ideology scores. Padó et al.[261] built a bipartite network of actors and claimed to understand the structures of political debates. Burfoot et al.[166] analyzed the sentiment of US congressional floor-debate transcripts with the help of document networks, where one speaker cites another was annotated. To predict the frames used in political discourse, Johnson et al.[262] assumed that politicians with shared ideologies are likely to frame issues in a similar way and retweet and/or follow each other on Twitter network. Rule et al.[48] studied the network of correlated words in textual corpora that span a long time, and identified that terms, concepts, and language use changes in American political consciousness since World War I in 1917.

For the election topic, Bovet and Makse[263] studied the dynamics and influence of fake news on Twitter during the 2016 US presidential election by analyzing the retweet networks formed by the top 100 news spreaders of different media categories. Grinberg et al.[81] studied the same problem with the help of a co-exposure network, where nodes are news websites and edges are shared-audience relationships. Bovet et al.[49] inferred the opinion of Twitter users in the context of the 2016 US presidential election based on both social network and hashtag co-occurrence network. Volkova et al.[264] also inferred user's political preferences between democrat and republican based on the Twitter social graph.

In the domain of **environment**, the network types used in different work are quite diverse. However, simple statistics and network centrality coefficients are still the most popular techniques for network analysis.

Farrell[76] constructed an organization network and concluded that organizations with corporate funding are more likely to write and spread texts that lead to ideological polarization on the climate change issue. Farrell[73] built a bipartite graph of the climate contrarian network with 4556 individuals and 164 contrarian organizations, in order to uncover the institutional and corporate structure of the climate change counter-movement. Barnes et al.[265] studied the information-sharing networks among tuna fishers to reveal how these social networks affect the incidental catch of sharks, a global environmental issue. Reino et al.[96] studied the global trade network of wild-caught birds with network centrality to analyze the bird invasion risk of different regions. Cámara-Leret et al.[266] proposed indigenous knowledge networks to describe the wisdom of indigenous people on plant species and the services they provide. Zheng et al.[267] and Hsieh et al.[268] utilized air quality monitoring data, human mobility, road network structures, and other information to suggest the best locations of new monitoring stations.

In the domain of **communication**, most works discuss the phenomenon of information diffusion (e.g., the spread of ideas, opinions, or products) in online or offline social networks.

For example, Guille and Hacid[269] employed feature engineering and a simple probabilistic model to characterize the temporal dynamics of information diffusion in social networks. Gómez-Gardeñes et al.[270] also studied the spread of social phenomena, such as behaviours, ideas, or products in the contact network of individuals. Pei et al.[99] and Zhang et al.[271] utilized various network centrality coefficients to detect the most influential information spreaders in online social networks. Brady et al.[64] analyzed the retweeting behaviours in online social networks and found that the expression of moral emotion is key for the spread of moral and political ideas. Gao et al.[272] compared the contact networks of users under emergency events and non-emergency events, in order to figure out how human communications will affect the propagation of situational awareness.

Some of these works especially focus on the spread of rumor, misinformation, and fake news, Quattrociocchi et al.[273] studied opinion dynamics on the network containing the interactions between gossipers, the influence network between gossiper and media, and the leader-follower relationship between media. Bovet and Makse[263] studied the dynamics and influence of fake news on Twitter during the 2016 US presidential election by analyzing the retweet networks formed by the top 100 news spreaders of different media categories. Shao et al.[274] studied the spread of low-credibility content in a retweet network, and concluded that social bots play an important role in spreading articles from low-credibility sources. In contrast, Vosoughi et al.[275] also studied the spread of false news on Twitter networks, and found that robots accelerate the spread of true and false news at the same rate, indicating that false news spreads faster because of human.

Besides, there is some work crossed with other domains. To find out whether restricting mobility or spreading disease prevention information better helps the control of diseases, Lima et al.[276] modeled human mobility and communications by an interconnected multiplex structure, where each node represents the population in a geographic area, and extended the model with a social network where relevant disease prevention information spreads. Luo et al.[277] measured individuals' location and influence in the social network from mobile and residential communication data, and found that an individual's location is highly correlated with personal economic status. Gomez and Lazer[278] studied how the distributions of knowledge and ability within a network of collective problem solvers contribute to the performance of the entire group. Farrell[73] built a bipartite graph of the climate contrarian network with 4556 individuals and 164 contrarian organizations, in order to uncover the institutional and corporate structure of the climate change counter-movement.

A.2.2 Embedding-based representation

In the domain of **sociology** , matrix factorization and topic models are widely used for learning user embeddings in a user-user network or user-item interaction network in the early 2010s. Recently, deep learning models, such as graph neural networks, are becoming the mainstream to encode structural information.

Many studies focus on the completion task, such as inferring the missing attributes or recommending potential friends/items. Kosinski et al.[114] applied singular value decomposition to the user-like matrix for learning users' embeddings, which were further utilized for predicting private traits. Pan et al.[184] utilized graph convolutional network to embed the users in a user network for occupation prediction. Yang et al.[112] applied matrix factorization to the social network including user-user friendship network and bipartite user-item interaction network for learning user and item embeddings, which were employed for friend and item recommendations. Fan et al.[118] and Wu et al.[119] used graph neural networks for social recommendation. Besides graph neural networks, Tang et al.[279] also employed LSTM for temporal modeling.

Other works can be formalized as sequential prediction, binary classification, and clustering, respectively. Li et al.[280] sampled diffusion sequences of users from the diffusion network, and applied RNN to encode the sequences and predict future users that would be influenced. Qiu et al.[281] employed graph neural networks to encode the ego network of each user, and then used the embeddings to classify whether the user will be influenced during information diffusion. Lu and Li[282] used graph attention network on user network for fake news detection. Zhong et al.[283] applied graph convolutional network on reply relationship network for controversy detection. Sachan et al.[113] employed a topic model to compute the community distribution of each user in a social network.

In the domain of **geography** , embedding-based representations of road networks are most widely used.

For example, Wang et al.[284] employed deep learning models, including recurrent neural networks, attention mechanism, and graph neural networks, to project each road into embedding-based representations and characterize the dynamics of traffic flow in road networks. Deng et al.[285] learned embeddings for each road in a road network via non-negative matrix factorization. The embeddings can encode both topological and temporal properties for traffic prediction. Li et al.[286] focused on travel time estimation in a road network and employed a multi-task learning framework to encode links and spatial-temporal factors. Sun et al.[287] developed a spatial-temporal latent factor model to identify the latent travel patterns and demands of urban region visitors. Pan et al.[288] used graph attention network to encode road networks, and RNN to further embed the temporal sequence of traffics for urban traffic prediction.

Besides, social media and social networks related to geo-locations are also explored. Zhang et al.[117] constructed a heterogeneous network with three types of nodes, i.e., location, time, and text, from Geo-Tagged Social Media (GTSM) data. Then they jointly encoded all spatial, temporal, and textual units into the same embedding space to capture the correlations for modeling people's activities in the urban space. Miura et al.[186] applied attention mechanism to user mention network extracted from Twitter to enhance the performance of geolocation prediction. Yang et al.[116] characterized a location-based social network, containing both user mobility data and the corresponding social network as a hypergraph, where a friendship is represented by an edge between two user nodes, and check-in is represented by a hyperedge among four nodes (a user, an activity type, a timestamp, and a POI). Network embedding methods are then employed for both friendship and location predictions.

For others, Yuan et al.[289] constructed a region transition network by connecting origin and destination regions of human mobility, and used the topic model to learn functional topic distributions for each region. Wang et al.[290] developed a driving state transition graph to characterize time-varying driving behaviour sequence, where nodes denote driving states (e.g., acceleration, turning right, etc.), and the weights of edges can be the frequency of state changes or the duration of state changes between two driving states. Then they employed a deep autoencoder to transform graphs into low-dimensional vectors and utilized RNN to incorporate temporal patterns.

In the domain of **economics** , a recent work[291] employed an attributed heterogeneous information network to characterize the behaviours and relationships between users, merchants, and devices. Then they used a fully neural-based model to model the representations of users for default probability prediction.

In the domain of **politics**, Stefanov et al.[192] applied node2vec[34] to a user-to-hashtag graph and a user-to-mention graph to learn users' embeddings, which can help better predict the stance and political leaning of media. Li and Goldwasser[191] employed GCN[36] to embed the social information graph, as well as text features, for identifying the political perspective of news media.

In the domain of **environment** , Shang et al.[292] estimated traffic conditions in a road network by filling the missing entries in an affinity matrix, where time slot embeddings, road embeddings, and feature embeddings are learned by matrix factorization.

In the domain of **communications**, Tang et al.[210] aimed to find the most influential users in a network on a specific topic, and how the influential users connect with each other. They characterized each user with topic distributions learned by a topic model, which can be seen as a non-negative embedding for each user. Distribution of these applications are listed in Table A1.

Acknowledgment

This work was supported by the National Key Research and Development Program of China (No. 2020AAA0106501), the National Natural Science Foundation of China (No. 62002029), and Beijing Academy of Artificial Intelligence (BAAI).

Table A1 Distribution of CSS applications.

Domain	Symbol		Embedding	
	Text	Network	Text	Network
Sociology	[39, 45, 49, 53, 57, 60, 62, 64–67, 86, 129, 130, 133–136, 138–152, 181, 293–314]	[39, 45, 49, 64, 81, 82, 85, 87, 89–93, 95, 100–105, 107–111, 136, 138, 141, 142, 144, 145, 150, 211–225, 227, 235, 237, 248, 253, 260, 265, 269, 270, 274, 275, 277, 278, 296, 298–301, 311, 313–406]	[62, 68–71, 113, 117, 134, 137, 139, 148, 152, 174–184, 187, 192, 193, 196, 279, 281, 283, 294, 407–435]	[112–115, 117–119, 184, 192, 279–283, 415]
Anthropology	[59, 173]	[59, 84, 85, 94, 97, 226–229]	[78]	
Psychology	[42, 51, 57, 63, 77, 158, 159, 436]	[83, 233–237]	[51, 77, 80, 160, 198–200]	[200]
Politics	[40, 48, 49, 56, 64, 67, 165, 166, 194, 262, 437–439]	[48, 49, 64, 76, 81, 166, 260–264, 439]	[76, 165, 191, 192, 194–197, 440, 441]	[191, 192]
Economics	[43, 169]	[241, 255–259, 442–451]	[75, 169, 193, 203–206]	[291]
Linguistics	[10, 37, 38, 47, 54–56, 61, 62, 153–157, 452, 453]	[230–232]	[44, 62, 72, 74, 207–209]	
Communication	[41, 52, 64, 167, 168],	[64, 73, 99, 263, 269–278]	[41, 73, 195, 210]	[210]
Geography	[161–164]	[88, 106, 111, 238–252, 254, 454–458]	[117, 185–190, 459]	[116, 117, 186, 284–290]
Environment	[58, 172]	[73, 76, 96, 265–268, 460, 461]	[73, 202]	[292]

References

[1] D. M. J. Lazer, A. Pentland, D. J. Watts, S. Aral, S. Athey, N. Contractor, D. Freelon, S. Gonzalez-Bailon, G. King, H. Margetts, et al., Computational social science: Obstacles and opportunities, *Science*, vol. 369, no. 6507, pp. 1060–1062, 2020.

[2] J. Zhang, W. Wang, F. Xia, Y. R. Lin, and H. H. Tong, Data-driven computational social science: A survey, *Big Data Res.*, vol. 21, p. 100145, 2020.

[3] D. Lazer, A. Pentland, L. Adamic, S. Aral, A. L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al., Computational social science, *Science*, vol. 323, no. 5915, pp. 721–723, 2009.

[4] M. Nadhom and P. Loskot, Survey of public data sources on the Internet usage and other Internet statistics, *Data Brief*, vol. 18, pp. 1914–1929, 2018.

[5] Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[6] Z. S. Harris, Distributional structure, *WORD* , vol. 10, no. 2&3, pp. 146–162, 1954.

[7] C. E. Shannon, A mathematical theory of communication, *Bell Syst. Tech. J.* , vol. 27, no. 3, pp. 379–423, 1948.

[8] G. Salton, E. A Fox, and H. Wu, Extended Boolean information retrieval, *Commun. ACM* , vol. 26, no. 11, pp. 1022–1036, 1983.

[9] J. W. Pennebaker, M. E. Francis, and R. J. Booth, *Linguistic Inquiry and Word Count*: *LIWC 2001*. Mahway, NJ, USA: Lawrence Erlbaum Associates, 2001.

[10] P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, et al., Human language reveals a universal positivity bias, *Proc. Natl. Acad. Sci. USA*, vol. 112, no. 8, pp. 2389–2394, 2015.

[11] A. Lenci, Distributional models of word meaning, *Ann. Rev. Linguist.*, vol. 4, pp. 151–171, 2018.

[12] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, Indexing by latent semantic analysis, *J. Am. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[13] J. Pennington, R. Socher, and C. D. Manning, Glove: Global vectors for word representation, in *Proc. 2014 Conf. on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 1532−1543.

[14] M. Sahlgren, An introduction to random indexing, in *Proc. Methods and Applications of Semantic Indexing Workshop at the 7th Int. Conf. on Terminology and Knowledge Engineering*, doi: 10.1111/j.1749-6632.1996.tb21128.x.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, Distributed representations of words and phrases and their compositionality, in *Proc. 27th Ann. Conf. on Neural Information Processing Systems 2013*, Lake Tahoe, NV, USA, 2013, pp. 3111−3119.

[16] F. Morin and Y. Bengio, Hierarchical probabilistic neural network language model, in *Proc. 10th Int. Workshop on Artificial Intelligence and Statistics*, Barbados, 2005, pp. 246−252.

[17] A. Mnih and Y. W. Teh, A fast and simple algorithm for training neural probabilistic language models, in *Proc. 29th Int. Conf. on Machine Learning*, Edinburgh, UK, 2012, pp. 1751−1759.

[18] O. Levy, Y. Goldberg, and I. Dagan, Improving distributional similarity with lessons learned from word Embeddings, *Trans. Assoc. Comput. Linguistics* , vol. 3, pp. 211–225, 2015.

[19] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet allocation, *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[20] M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand, Structural topic models for open-ended survey responses, *Am. J. Polit. Sci.* , vol. 58, no. 4, pp. 1064–1082, 2014.

[21] A. Baddeley, Working memory, *Science* , vol. 255, no. 5044, pp. 556–559, 1992.

[22] W. James, F Burkhardt, F. Bowers, and I. K. Skrupskelis, *The Principles of Psychology*. London, UK: Macmillan, 1890.

[23] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, A convolutional neural network for modelling sentences, in *Proc. 52nd Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Baltimore, MD, USA, 2014, pp. 655−665.

[24] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, Recurrent neural network based language model, in *Proc. 11th Ann. Conf. of the Int. Speech Communication Association*, Makuhari, Japan, 2010, pp. 1045−1048.

[25] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, in *Proc. 2014 Conf. on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 1724−1734.

[26] S. Hochreiter and J. Schmidhuber, Long short-term memory, *Neural Comput.* , vol. 9, no. 8, pp. 1735–1780, 1997.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, in *Proc. Ann. Conf. on Neural Information Processing Systems 2017*, Long Beach, CA, USA, 2017, pp. 5998−6008.

[28] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in *Proc. Conf. of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, *Volume 1* (*Long and Short Papers*), Minneapolis, MN, USA, 2019, pp. 4171−4186.

[29] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, *et al.*, Languagemodels are few-shot learners, in *Proc. of 34th Annual Conference on Neural Information Processing Systems*, https://www.researchgate.net/publication/341724146_Language_Models_are_Few-Shot_Learners, 2020.

[30] B. Perozzi, R. Al-Rfou, and S. Skiena, DeepWalk: Online learning of social representations, in *Proc. 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data*

Mining, New York City, NY, USA, 2014, pp. 701-710.

[31] M. Belkin and P. Niyogi, Laplacian Eigenmaps and spectral techniques for embedding and clustering, in *Proc. Neural Information Processing Systems*: *Natural and Synthetic*, Vancouver, Canada, 2001, pp. 585-591.

[32] F. Fouss, A. Pirotte, J. M. Renders, and M. Saerens, Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation, *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 3, pp. 355–369, 2007.

[33] R. Andersen, F. R. K. Chung, and K. J. Lang, Local graph partitioning using Pagerank vectors, in *Proc. 47th Ann. IEEE Symp. on Foundations of Computer Science*, Berkeley, CA, USA, 2006, pp. 475−486.

[34] A. Grover and J. Leskovec, Node2Vec: Scalable feature learning for networks, in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 855-864.

[35] J. Tang, M. Qu, M. Z. Wang, M. Zhang, J. Yan, and Q. Z. Mei, LINE: Large-scale information network embedding, in *Proc. 24th Int. Conf. on World Wide Web*, Florence, Italy, 2015, pp. 1067-1077.

[36] T. N. Kipf and M. Welling, Semi-supervised classification with graph convolutional networks, in *Proc. 5th Int. Conf. on Learning Representations*, Toulon, France, 2017.

[37] C. Yang, Ontogeny and phylogeny of language, *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 16, pp. 6324–6327, 2013.

[38] J. B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, *et al*., Quantitative analysis of culture using millions of digitized books, *Science*, vol. 331, no. 6014, pp. 176−182, 2011.

[39] E. E. Bruch and M. E. J. Newman, Aspirational pursuit of mates in online dating markets, *Sci. Adv.*, vol. 4, no. 8, p. eaap9815, 2018.

[40] A. Lupia, S. Soroka, and A. Beatty, What does congress want from the National Science Foundation? A content analysis of remarks from 1995 to 2018 *Sci. Adv.*, vol. 6, no. 33, p. eaaz6300, 2020.

[41] K. Sheshadri and M. P. Singh, The public and legislative impact of hyperconcentrated topic news, *Sci. Adv.*, vol. 5, no. 8, p. eaat8296, 2019.

[42] S. A. Golder and M. W. Macy, Diurnal and seasonal mood vary with work, sleep, and Daylength across diverse cultures, *Science*, vol. 333, no. 6051, pp. 1878–1881, 2011.

[43] M. Alanyali, H. S. Moat, and T. Preis, Quantifying the relationship between financial news and the stock market, *Sci. Rep.*, vol. 3, no. 1, p. 3578, 2013.

[44] A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, Natural speech reveals the semantic maps that tile human cerebral cortex, *Nature*, vol. 532, no. 7600, pp. 453–458, 2016.

[45] M. Stella, E. Ferrara, and M. De Domenico, Bots increase exposure to negative and inflammatory content in online social systems, *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 49, pp. 12435–12440, 2018.

[46] C. Ramiro, M. Srinivasan, B. C. Malt, and Y. Xu, Algorithms in the historical emergence of word senses, *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 10, pp. 2323–2328, 2018.

[47] J. C. Jackson, J. Watts, T. R. Henry, J. M. List, R. Forkel, P. J. Mucha, S. J. Greenhill, R. D. Gray, and K. A. Lindquist, Emotion semantics show both cultural variation and universal structure, *Science*, vol. 366, no. 6472, pp. 1517–1522, 2019.

[48] A. Rule, J. P. Cointet, and P. S. Bearman, Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790-2014, *Proc. Natl. Acad. Sci. USA*, vol. 112, no. 35, pp. 10837–10844, 2015.

[49] A. Bovet, F. Morone, and H. A. Makse, Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald trump, *Sci. Rep.*, vol. 8, no. 1, p. 8673, 2018.

[50] D. T. Citron and P. Ginsparg, Patterns of text reuse in a scientific corpus, *Proc. Natl. Acad. Sci. USA*, vol. 112, no. 1, pp. 25–30, 2015.

[51] J. C. Eichstaedt, R. J. Smith, R. M. Merchant, L. H. Ungar, P. Crutchley, D. Preoţuc-Pietro, D. A. Asch, and H. A. Schwartz, Facebook language predicts depression in medical records, *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 44, pp. 11203–11208, 2018.

[52] J. Green, J. Edgerton, D. Naftel, K. Shoub, and S. J. Cranmer, Elusive consensus: Polarization in elite communication on the COVID-19 pandemic, *Sci. Adv.*, vol. 6, no. 28, p. eabc2717, 2020.

[53] E. Bakshy, S. Messing, and L. A. Adamic, Exposure to ideologically diverse news and opinion on Facebook, *Science*, vol. 348, no. 6239, pp. 1130–1132, 2015.

[54] S. T. Piantadosi, H. Tily, and E. Gibson, Word lengths are optimized for efficient communication, *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 9, pp. 3526–3529, 2011.

[55] R. Futrell, K. Mahowald, and E. Gibson, Large-scale evidence of dependency length minimization in 37 languages, *Proc. Natl. Acad. Sci. USA*, vol. 112, no. 33, pp. 10336–10341, 2015.

[56] K. N. Jordan, J. Sterling, J. W. Pennebaker, and R. L. Boyd, Examining long-term trends in politics and culture through language of political leaders and cultural institutions, *Proc. Natl. Acad. Sci. USA*, vol. 116, no. 9, pp. 3476–3481, 2019.

[57] M. R. Frank, L. Mitchell, P. S. Dodds, and C. M. Danforth, Happiness and the patterns of life: A study of Geolocated tweets, *Sci. Rep.*, vol. 3, no. 1, p. 2625, 2013.

[58] Y. Kryvasheyeu, H. H. Chen, N. Obradovich, E. Moro, P. V. Hentenryck, J. Fowler, and M. Cebrian, Rapid assessment of disaster damage using social media activity, *Sci. Adv.*, vol. 2, no. 3, p. e1500779, 2016.

[59] S. Klingenstein, T. Hitchcock, and S. DeDeo, The civilizing process in London's Old Bailey, *Proc. Natl. Acad. Sci. USA*, vol. 111, no. 26, pp. 9419–9424, 2014.

[60] C. Catalini, N. Lacetera, and A. Oettl, The incidence and role of negative citations in science, *Proc. Natl. Acad. Sci. USA*, vol. 112, no. 45, pp. 13823–13826, 2015.

[61] R. L. Boyd, K. G. Blackburn, and J. W. Pennebaker, The narrative arc: Revealing core narrative structures through text analysis, *Sci. Adv.*, vol. 6, no. 32, p. eaba2196, 2020.

[62] J. M. Hughes, N. J. Foti, D. C. Krakauer, and D. N. Rockmore, Quantitative patterns of stylistic influence in the evolution of literature, *Proc. Natl. Acad. Sci. USA*, vol. 109, no. 20, pp. 7682–7686, 2012.

[63] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock,

Experimental evidence of massive-scale emotional contagion through social networks, *Proc. Natl. Acad. Sci. USA*, vol. 111, no. 24, pp. 8788–8790, 2014.

[64] W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, and J. J. Van Bavel, Emotion shapes the diffusion of moralized content in social networks, *Proc. Natl. Acad. Sci. USA*, vol. 114, no. 28, pp. 7313–7318, 2017.

[65] N. M. Jones, R. R. Thompson, C. D. Schetter, and R. C. Silver, Distress and rumor exposure on social media during a campus lockdown, *Proc. Natl. Acad. Sci. USA*, vol. 114, no. 44, pp. 11663–11668, 2017.

[66] M. Del Vicario, G. Vivaldo, A. Bessi, F. Zollo, A. Scala, G. Caldarelli, and W. Quattrociocchi, Echo chambers: Emotional contagion and group polarization on facebook, *Sci. Rep.*, vol. 6, no. 1, p. 37825, 2016.

[67] M. Alizadeh, J. N. Shapiro, C. Buntain, and J. A. Tucker, Content-based features predict social media influence operations, *Sci. Adv.*, vol. 6, no. 30, p. eabb5824, 2020.

[68] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes, *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 16, pp. E3635–E3644, 2018.

[69] A. Caliskan, J. J. Bryson, and A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, *Science* , vol. 356, no. 6334, pp. 183–186, 2017.

[70] E. Sivak and I. Smirnov, Parents mention sons more often than daughters on social media, *Proc. Natl. Acad. Sci. USA*, vol. 116, no. 6, pp. 2039–2041, 2019.

[71] A. Gerow, Y. N. Hu, J. Boyd-Graber, D. M. Blei, and J. A. Evans, Measuring discursive influence across scholarship, *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 13, pp. 3308–3313, 2018.

[72] E. Bokányi, D. Kondor, L. Dobos, T. Sebök, J. Stéger, I. Csabai, and G. Vattay, Race, religion and the city: Twitter word frequency patterns reveal dominant demographic dimensions in the United States, *Palgrave Commun.*, vol. 2, no. 1, p. 16010, 2016.

[73] J. Farrell, Network structure and influence of the climate change counter-movement, *Nat. Climate Change* , vol. 6, no. 4, pp. 370–374, 2016.

[74] B. C. Roy, M. C. Frank, P. DeCamp, M. Miller, and D. Roy, Predicting the birth of a spoken word, *Proc. Natl. Acad. Sci. USA*, vol. 112, no. 41, pp. 12663–12668, 2015.

[75] C. Curme, T. Preis, H. E. Stanley, and H. S. Moat, Quantifying the semantics of search behavior before stock market moves, *Proc. Natl. Acad. Sci. USA*, vol. 111, no. 32, pp. 11600–11605, 2014.

[76] J. Farrell, Corporate funding and ideological polarization about climate change, *Proc. Natl. Acad. Sci. USA*, vol. 113, no. 1, pp. 92–97, 2016.

[77] K. Jaidka, S. Giorgi, H. A. Schwartz, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods, *Proc. Natl. Acad. Sci. USA*, vol. 117, no. 19, pp. 10165–10171, 2020.

[78] E. Fetaya, Y. Lifshitz, E. Aaron, and S. Gordin, Restoration of fragmentary Babylonian texts using recurrent neural networks, *Proc. Natl. Acad. Sci. USA*, vol. 117, no. 37, pp. 22743–22751, 2020.

[79] M. Hahn, D. Jurafsky, and R. Futrell, Universals of word order reflect optimization of grammars for efficient communication, *Proc. Natl. Acad. Sci. USA* , vol. 117, no. 5, pp. 2347–2353, 2020.

[80] M. Mooijman, J. Hoover, Y. Lin, H. Ji, and M. Dehghani, Moralization in social networks and the emergence of violence during protests, *Nat. Hum. Behav.*, vol. 2, no. 6, pp. 389–396, 2018.

[81] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer, Fake news on Twitter during the 2016 U.S. presidential election, *Science* , vol. 363, no. 6425, pp. 374–378, 2019.

[82] M. X. Li, Z. Q. Jiang, W. J. Xie, S. Miccichè, M. Tumminello, W. X. Zhou, and R. N. Mantegna, A comparative analysis of the statistical properties of large mobile phone calling networks, *Sci. Rep.*, vol. 4, no. 1, p. 5132, 2014.

[83] K. M. Turetsky, V. Purdie-Greenaway, J. E. Cook, J. P. Curley, and G. L. Cohen, A psychological intervention strengthens students' peer social networks and promotes persistence in STEM, *Sci. Adv.* , vol. 6, no. 45, p. eaba9221, 2020.

[84] C. L. Apicella, F. W. Marlowe, J. H. Fowler, and N. A. Christakis, Social networks and cooperation in hunter-gatherers, *Nature*, vol. 481, no. 7382, pp. 497–501, 2012.

[85] J. D. Boardman, B. W. Domingue, and J. M. Fletcher, How social and genetic factors predict friendship networks, *Proc. Natl. Acad. Sci. USA* , vol. 109, no. 43, pp. 17377–17381, 2012.

[86] A. Clauset, S. Arbesman, and D. B. Larremore, Systematic inequality and hierarchy in faculty hiring networks, *Sci. Adv.*, vol. 1, no. 1, p. e1400005, 2015.

[87] C. M. Trujillo and T. M. Long, Document co- citation analysis to enhance transdisciplinary research, *Sci. Adv.*, vol. 4, no. 1, p. e1701130, 2018.

[88] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. O. Buckee, Quantifying the impact of human mobility on malaria, *Science*, vol. 338, no. 6104, pp. 267–270, 2012.

[89] H. H. Jo, J. Saramäki, R. I. M. Dunbar, and K. Kaski, Spatial patterns of close relationships across the lifespan, *Sci. Rep.*, vol. 4, no. 1, p. 6988, 2014.

[90] C. Parkinson, A. M. Kleinbaum, and T. Wheatley, Similar neural responses predict friendship, *Nat. Commun.*, vol. 9, no. 1, p. 332, 2018.

[91] L. F. Wu, D. S. Wang, and J. A. Evans, Large teams develop and small teams disrupt science and technology, *Nature*, vol. 566, no. 7744, pp. 378–382, 2019.

[92] P. S. Park, J. E. Blumenstock, and M. W. Macy, The strength of long-range ties in population-scale social networks, *Science* , vol. 362, no. 6421, pp. 1410–1413, 2018.

[93] D. Garcia, Leaking privacy and shadow profiles in online social networks, *Sci. Adv.* , vol. 3, no. 8, p. e1701172, 2017.

[94] M. Schich, C. M. Song, Y. Y. Ahn, A. Mirsky, M. Martino, A. L. Barabási, and D. Helbing, A network framework of cultural history, *Science* , vol. 345, no. 6196, pp. 558–562, 2014.

[95] P. Manrique, Z. F. Cao, A. Gabriel, J. Horgan, P. Gill, H. Qi, E. M. Restrepo, D. Johnson, S. Wuchty, C. M. Song, et al., Women's connectivity in extreme networks, *Sci. Adv.*, vol. 2, no. 6, p. e1501742, 2016.

[96] L. Reino, R. Figueira, P. Beja, M. B. Araújo, C. Capinha,

and D. Strubbe, Networks of global bird invasion altered by regional trade ban, *Sci. Adv.* , vol. 3, no. 11, p. e1700783, 2017.

[97] J. P. Hart, J. Birch, and C. G. St-Pierre, Effects of population dispersal on regional signaling networks: An example from northern Iroquoia, *Sci. Adv.*, vol. 3, no. 8, p. e1700497, 2017.

[98] S. P. Fraiberger, R. Sinatra, M. Resch, C. Riedl, and A. L. Barabási, Quantifying reputation and success in art, *Science*, vol. 362, no. 6416, pp. 825–829, 2018.

[99] S. Pei, L. Muchnik, J. S. J. Andrade Jr, Z. M. Zheng, and H. A. Makse, Searching for superspreaders of information in real-world social media, *Sci. Rep.*, vol. 4, no. 1, p. 5547, 2014.

[100] M. Waniek, T. P. Michalak, M. J. Wooldridge, and T. Rahwan, Hiding individuals and communities in a social network, *Nat. Hum. Behav.* , vol. 2, no. 2, pp. 139–147, 2018.

[101] M. M. Dankulov, R. Melnik, and B. Tadić, The dynamics of meaningful social interactions and the emergence of collective knowledge, *Sci. Rep.* , vol. 5, no. 1, p. 12197, 2015.

[102] A. Asikainen, G. Iñiguez, J. Ureña-Carrión, K. Kaski, and M. Kivelä, Cumulative effects of triadic closure and homophily in social networks, *Sci. Adv.* , vol. 6, no. 19, p. eaax7310, 2020.

[103] B. Charoenwong, A. Kwan, and V. Pursiainen, Social connections with COVID-19-affected areas increase compliance with mobility restrictions, *Sci. Adv.* , vol. 6, no. 47, p. eabc3054, 2020.

[104] S. Aral and D. Walker, Identifying influential and susceptible members of social networks, *Science*, vol. 337, no. 6092, pp. 337–341, 2012.

[105] Y. H. Eom and H. H. Jo, Generalized friendship paradox in complex networks: The case of scientific collaboration, *Sci. Rep.*, vol. 4, no. 1, p. 4603, 2014.

[106] A. A. Ganin, M. Kitsak, D. Marchese, J. M. Keisler, T. Seager, and I. Linkov, Resilience and efficiency in transportation networks, *Sci. Adv.* , vol. 3, no. 12, p. e1701079, 2017.

[107] F. A. Massucci, J. Wheeler, R. Beltrán-Debón, J. Joven, M. Sales-Pardo, and R. Guimerà, Inferring propagation paths for sparsely observed perturbations on complex networks, *Sci. Adv.*, vol. 2, no. 10, p. e1501638, 2016.

[108] X. Teng, S. Pei, F. Morone, and H. A. Makse, Collective influence of multiple spreaders evaluated by tracing real information flow in large-scale social networks, *Sci. Rep.*, vol. 6, no. 1, p. 36043, 2016.

[109] M. Medo, M. S. Mariani, A. Zeng, and Y. C. Zhang, Identification and impact of discoverers in online social systems, *Sci. Rep.*, vol. 6, no. 1, p. 34218, 2016.

[110] L. Kovanen, K. Kaski, J. Kertész, and J. Saramäki, Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences, *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 45, pp. 18070–18075, 2013.

[111] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte, Uncovering space-independent communities in spatial networks, *Proc. Natl. Acad. Sci. USA* , vol. 108, no. 19, pp. 7663–7668, 2011.

[112] S. H. Yang, B. Long, A. Smola, S. H. Yang, B. Long, A. Smola, N. Sadagopan, Z. H. Zheng, and H. Zha, Like like alike: Joint friendship and interest propagation in social networks, in *Proc. 20th Int. Conf. on World Wide Web*, Hyderabad, India, 2011, pp. 537−546.

[113] M. Sachan, D. Contractor, M. Sachan, D. Contractor, T. A. Faruquie, and L. V. Subramaniam, Using content and interactions for discovering communities in social networks, in *Proc. 21st Int. Conf. on World Wide Web*, Lyon, France, 2012, pp. 331−340.

[114] M. Kosinski, D. Stillwell, and T. Graepel, Private traits and attributes are predictable from digital records of human behavior, *Proc. Natl. Acad. Sci. USA* , vol. 110, no. 15, pp. 5802–5805, 2013.

[115] G. Wang, Y. C. Zhao, X. X. Shi, and P. S. Yu, Magnet community identification on social networks, in *Proc. 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Beijing, China, 2012, pp. 588−596.

[116] D. Q. Yang, B. Q. Qu, J. Yang, and P. Cudre-Mauroux, Revisiting user mobility and social relationships in LBSNs: A hypergraph embedding approach, in *Proc. World Wide Web Conf.*, San Francisco, CA, USA, 2019, pp. 2147−2157.

[117] C. Zhang, K. Y. Zhang, Q. Yuan, H. R. Peng, Y. Zheng, T. Hanratty, S. W. Wang, and J. W. Han, Regions, periods, activities: Uncovering urban dynamics via cross-modal representation learning, in *Proc. 26th Int. Conf. on World Wide Web*, Perth, Australia, 2017, pp. 361−370.

[118] W. Q. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. L. Tang, and D. W. Yin, Graph neural networks for social recommendation, in *Proc. World Wide Web Conf.*, San Francisco, CA, USA, 2019, pp. 417−426.

[119] Q. T. Wu, H. R. Zhang, X. F. Gao, P. He, P. Weng, H. Gao, and G. H. Chen, Dual graph attention networks for deep latent representation of multifaceted social effects in recommender systems, in *Proc. World Wide Web Conf.*, San Francisco, CA, USA, 2019, pp. 2091−2102.

[120] C. W. J. Granger, Investigating causal relations by econometric models and cross-spectral methods, *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[121] Q. S. Zhang and S. C. Zhu, Visual interpretability for deep learning: A survey, *Front. Inf. Technol. Electr. Eng.*, vol. 19, no. 1, pp. 27–39, 2018.

[122] Y. Belinkov and J. Glass, Analysis methods in neural language processing: A survey, *Trans. Assoc. Comput. Linguistics*, vol. 7, pp. 49–72, 2019.

[123] M. D, Zeiler and R. Fergus, Visualizing and understanding convolutional networks, in *Proc. 13th European Conf. on Computer Vision*, Zurich, Switzerland, 2014, pp. 818−833.

[124] Y. H. Liu, M. Ott, N. Goyal, J. F. Du, M. Joshi, D. Q. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, RoBERta: A robustly optimized BERT pretraining approach, arXiv preprint arXiv: 1907.11692, 2019.

[125] J. Zhou, G. Q. Cui, S. D. Hu, Z. Y. Zhang, C. Yang, Z. Y. Liu, L. F. Wang, C. C. Li, and M. S. Sun, Graph neural networks: A review of methods and applications, *AI Open*, vol. 1, pp. 57–81, 2020.

[126] L. F. Wu, Y. Chen, H. Ji, and Y. Y. Li, Deep learning on graphs for natural language processing, in *Proc. 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials* , Seattle, WA, USA, 2021, pp.

11−14.

[127] R. Ying, R. N. He, K. F. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, Graph convolutional neural networks for web-scale recommender systems, in *Proc. 24th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, London, UK, 2018, pp. 974-983.

[128] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Inf. Fusion*, vol. 58, pp. 82–115, 2020.

[129] C. Boididou, S. Papadopoulos, Y. Kompatsiaris, S. Schifferes, and N. Newman, Challenges of computational verification in social multimedia, in *Proc. 23rd Int. Conf. on World Wide Web*, Seoul, Republic of Korea, 2014, pp. 743−748.

[130] Z. Zhao, P. Resnick, and Q. Z. Mei, Enquiring minds: Early detection of rumors in social media from enquiry posts, in *Proc. 24th Int. Conf. on World Wide Web*, Florence, Italy, 2015, pp. 1395−1405.

[131] S. Kumar, R. West, and J. Leskovec, Disinformation on the web: Impact, characteristics, and detection of Wikipedia hoaxes, in *Proc. 25th Int. Conf. on World Wide Web*, Montreal, Canada, 2016, pp. 591−602.

[132] K. Popat, S. Mukherjee, J. Strötgen, and G. Weikum, Where the truth lies: Explaining the credibility of emerging claims on the web and social media, in *Proc. 26th Int. Conf. on World Wide Web Companion*, Perth, Australia, 2017, pp. 1003−1012.

[133] F. Yang, Y. Liu, X. H. Yu, and M. Yang, Automatic detection of rumor on Sina Weibo, in *Proc. ACM SIGKDD Workshop on Mining Data Semantics*, New York, NY, USA, 2012, p. 13.

[134] G. Bhatt, A. Sharma, S. Sharma, A. Sharma, A. Nagpal, B. Raman, and A. Mittal, Combining neural, statistical and external features for fake news stance identification, in *Proc. Proc. Web Conf. 2018*, Lyon, France, 2018, pp. 1353−1357.

[135] A. Gupta, H. Lamba, P. Kumaraguru, and A. Joshi, Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy, in *Proc. 22nd Int. Conf. on World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 729−736.

[136] L. Flekova, O. Ferschke, and I. Gurevych, What makes a good biography?: Multidimensional quality analysis based on Wikipedia article feedback data, in *Proc. 23rd Int. Conf. on World Wide Web*, Seoul, Republic of Korea, 2014, pp. 855−866.

[137] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, Deep learning for hate speech detection in tweets, in *Proc. 26th Int. Conf. on World Wide Web Companion*, Perth, Australia, 2017, pp. 759−760.

[138] Z. Savvas, B. Bradlyn, E. De Cristofaro, H. Kwak, M. Sirivianos, G. Stringhini, and J. Blackburn, What is gab: A bastion of free speech or an alt-right echo chamber, in *Proc. Web Conf. 2018*, Lyon, France, 2018, pp. 1007−1014.

[139] C. Nobata, J. R. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, Abusive language detection in online user content, in *Proc. 25th Int. Conf. on World Wide Web*, Montreal, Canada, 2016, pp. 145−153.

[140] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali, Measuring #Gamergate: A tale of hate, sexism, and bullying, in *Proc. 26th Int. Conf. on World Wide Web Companion*, Perth, Australia, 2017, pp. 1285−1290.

[141] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, BotOrNot: A system to evaluate social bots, in *Proc. 25th Int. Conf. Companion on World Wide Web*, Montreal, Canada, 2016, pp. 273−274.

[142] S. Kumar, J. Cheng, J. Leskovec, and V. S. Subrahmanian, An army of me: Sockpuppets in online discussion communities, in *Proc. 26th Int. Conf. on World Wide Web*, Perth, Australia, 2017, pp. 857−866.

[143] R. S. Portnoff, S. Afroz, G. Durrett, J. K. Kummerfeld, T. Berg-Kirkpatrick, D. McCoy, K. Levchenko, and V. Paxson, Tools for automated analysis of cybercriminal markets, in *Proc. 26th Int. Conf. on World Wide Web*, Perth, Australia, 2017, pp. 657−666.

[144] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira, Exploiting innocuous activity for correlating users across sites, in *Proc. 22nd Int. Conf. on World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 447−458.

[145] P. Jain, P. Kumaraguru, and A. Joshi, @i seek ‘fb. me’ identifying users across multiple online social networks, in *Proc. 22nd Int. Conf. on World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 1259−1268.

[146] M. Imran, S. Elbassuoni, C. Castillo, F. Diaz, and P. Meier, Practical extraction of disaster-relevant information from social media, in *Proc. 22nd Int. Conf. on World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 1021−1024.

[147] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, AIDR: Artificial intelligence for disaster response, in *Proc. 23rd Int. Conf. on World Wide Web*, Seoul, Republic of Korea, 2014, pp. 159−162.

[148] S. Cresci, M. Tesconi, A. Cimino, and F. Dell’Orletta, A linguistically-driven approach to cross-event damage assessment of natural disasters from social media messages, in *Proc. 24th Int. Conf. on World Wide Web*, Florence, Italy, 2015, pp. 1195−1200.

[149] P. Bramsen, M. Escobar-Molano, A. Patel, and R. Alonso, Extracting social power relationships from natural language, in *Proc. 49th Ann. Meeting of the Association for Computational Linguistics*: *Human Language Technologies-Volume 1*, Portland, OR, USA, 2011, pp. 773−782.

[150] L. Yang, T. Sun, M. Zhang, and Q. Z. Mei, We know what @you #tag: Does the dual role affect hashtag adoption? in *Proc. 21st Int. Conf. on World Wide Web*, Lyon, France, 2012, pp. 261−270.

[151] J. Lehmann, B. Gonçalves, J. J. Ramasco, and C. Cattuto, Dynamical classes of collective attention in twitter, in *Proc. 21st Int. Conf. on World Wide Web*, Lyon, France, 2012, pp. 251−260.

[152] B. Fu, J. L. Lin, L. Li, C. Faloutsos, J. I. Hong, and N. M. Sadeh, Why people hate your app. Making sense of user feedback in a mobile app store, in *Proc. 19th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Chicago, IL, USA, 2013, pp. 1276−1284.

[153] D. Hovy, A. Johannsen, and A. Søgaard, User review sites as a resource for large-scale sociolinguistic studies, in *Proc. 24th Int. Conf. on World Wide Web*, Florence, Italy, 2015, pp. 452−461.

[154] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts, A computational approach to politeness with application to social factors, in *Proc. 51st Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Sofia, Bulgaria, 2013, pp. 250−259.

[155] C. Danescu-Niculescu-Mizil, J. Cheng, J. Kleinberg, and L. Lee, You had me at hello: How phrasing affects memorability, in *Proc. 50th Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Jeju Island, Republic of Korea, 2012, pp. 892−901.

[156] C. H. Tan, L. Lee, and B. Pang, The effect of wording on message propagation: Topic- and author- controlled natural experiments on Twitter, in *Proc. 52nd Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Baltimore, MD, USA, 2014, pp. 175-185.

[157] C. Hube and B. Fetahu, Detecting biased statements in Wikipedia, in *Proc. Web Conf. 2018*, Lyon, France, 2018, pp. 1779−1786.

[158] X. T. Chen, M. D. Sykora, T. W. Jackson, and S. Elayan, What about mood swings: Identifying depression on twitter with temporal measures of emotions, in *Proc. Web Conf. 2018*, Lyon, France, 2018, pp. 1653−1660.

[159] J. P. Chang, J. Cheng, and C. Danescu-Niculescu-Mizil, Don't let me be misunderstood: Comparing intentions and perceptions in online discussions, in *Proc. Web Conf. 2020*, Taipei, China, 2020, pp. 2066−2077.

[160] M. L Kern, P. X. McCarthy, D. Chakrabarty, and M. A. Rizoiu, Social media-predicted personality traits and values can help match people to their ideal jobs, *Proc. Natl. Acad. Sci. USA*, vol. 116, no. 52, pp. 26459–26464, 2019.

[161] Y. Ikawa, M. Enoki, and M. Tatsubori, Location inference using microblog messages, in *Proc. 21st Int. Conf. on World Wide Web*, Lyon, France, 2012, pp. 687−690.

[162] K. M. Ryoo and S. Moon, Inferring twitter user locations with 10 km accuracy, in *Proc. 23rd Int. Conf. on World Wide Web*, Seoul, Republic of Korea, 2014, pp. 643−648.

[163] B. Wing and J. Baldridge, Simple supervised document geolocation with geodesic grids, in *Proc. 49th Ann. Meeting of the Association for Computational Linguistics*: *Human Language Technologies*, Portland, OR, USA, 2011, pp. 955−964.

[164] J. Kim, M. Cha, and T. Sandholm, Socroutes: Safe routes based on tweet sentiments, in *Proc. 23rd Int. Conf. on World Wide Web*, Seoul, Republic of Korea, 2014, pp. 179−182.

[165] D. Preoţiuc-Pietro, Y. Liu, D. Hopkins, and L. Ungar, Beyond binary labels: Political ideology prediction of twitter users, in *Proc. 55th Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Vancouver, Canada, 2017, pp. 729−740.

[166] C. Burfoot, S. Bird, and T. Baldwin, Collective classification of congressional floor-debate transcripts, in *Proc. 49th Ann. Meeting of the Association for Computational Linguistics*: *Human Language Technologies*, Portland, OR, USA, 2011, pp. 1506-1515.

[167] M. Jenders, G. Kasneci, and F. Naumann, Analyzing and predicting viral tweets, in *Proc. 22nd Int. Conf. on World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 657-664.

[168] I. Arous, J. Yang, M. Khayati, and P. Cudré-Mauroux, OpenCrowd: A human-AI collaborative approach for finding social influencers via open-ended answers aggregation, in *Proc. Web Conf. 2020* , Taipei, China, 2020, pp. 1851−1862.

[169] B. Y. Xie, R. J. Passonneau, L. Wu, and G. G. Creamer, Semantic frames to predict stock price movement, in *Proc. 51st Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Sofia, Bulgaria, 2013, pp. 873−883.

[170] A. Agarwal, B. Y. Xie, I. Vovsha, O. Rambow, and R. Passonneau, Sentiment analysis of twitter data, in *Proc. Workshop on Languages in Social Media*, Stroudsburg, PA, USA, 2011, pp. 30-38.

[171] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, Sentiment strength detection in short informal text, *J. Am. Soc. Inf. Sci. Technol.* , vol. 61, no. 12, pp. 2544–2558, 2010.

[172] S. Ghosh and M. S. Desarkar, Class specific TF-IDF boosting for short-text classification: Application to short-texts generated during disasters, in *Proc. Web Conf. 2018*, Lyon, France, 2018, pp. 1629−1637.

[173] S. A Mehr, M. Singh, D. Knox, D. M. Ketter, D. Pickens-Jones, S. Atwood, C. Lucas, N. Jacoby, A. A. Egner, E. J. Hopkins, et al., Universality and diversity in human song, *Science*, vol. 366, no. 6468, p. eaax0868, 2019.

[174] P. Singer, F. Lemmerich, R. West, L. Zia, E. Wulczyn, M. Strohmaier, and J. Leskovec, Why we read Wikipedia? in *Proc. 26th Int. Conf. on World Wide Web*, Perth, Australia, 2017, pp. 1591−1600.

[175] J. Weerasinghe, B. Flanigan, A. J. Stein, D. McCoy, and R. Greenstadt, The pod people: Understanding manipulation of social media popularity via reciprocity abuse, in *Proc. Web Conf. 2020*, Taipei, China, 2020, pp. 1874−1884.

[176] W. Y. Wang, E. Mayfield, S. Naidu, and J. Dittmar, Historical analysis of legal opinions with a sparse mixed-effects latent variable model, in *Proc. 50th Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Jeju Island, Republic of Korea, 2012, pp. 740−749.

[177] J. Ma, W. Gao, and K. F. Wong, Detect rumor and stance jointly by neural multi-task learning, in *Proc. Web Conf. 2018*, Lyon, France, 2018, pp. 585−593.

[178] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, MVAE: Multimodal variational autoencoder for fake news detection, in *Proc. World Wide Web Conf.*, San Francisco, CA, USA, 2019, pp. 2915−2921.

[179] J. Ma, W. Gao, and K. F. Wong, Detect rumors on twitter by promoting information campaigns with generative adversarial learning, in *Proc. World Wide Web Conf.*, San Francisco, CA, USA, 2019, pp. 3049−3055.

[180] Q. Zhang, A. Lipani, S. S. Liang, and E. Yilmaz, Reply-aided detection of misinformation via bayesian deep learning, in *Proc. World Wide Web Conf.*, San Francisco,

CA, USA, 2019, pp. 2333−2343.

[181] H. Almerekhi, H. Kwak, J. Salminen, and B. J. Jansen, Are these comments triggering? Predicting triggers of toxicity in online discussions, in *Proc. Web Conf. 2020*, Taipei, China, 2020, pp. 3033−3040.

[182] Z. J. Wang, S. A. Hale, D. I. Adelani, P. A. Grabowicz, T. Hartmann, F. Flöck, and D. Jurgens, Demographic inference and representative population estimates from multilingual social media data, in *Proc. World Wide Web Conf.*, San Francisco, CA, USA, 2019, pp. 2056−2067.

[183] S. Wilson and R. Mihalcea, Predicting human activities from user-generated content, in *Proc. 57th Ann. Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 2572−2582.

[184] J. Q. Pan, R. Bhardwaj, W. Lu, H. L. Chieu, X. H. Pan, and N. Y. Puay, Twitter homophily: Network based prediction of user's occupation, in *Proc. 57th Ann. Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 2633−2638.

[185] A. Ahmed, L. J. Hong, and A. J. Smola, Hierarchical geographical modeling of user locations from social media posts, in *Proc. 22nd Int. Conf. on World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 25−36.

[186] Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma, Unifying text, metadata, and user network representations with a neural network for Geolocation prediction, in *Proc. 55th Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Vancouver, Canada, 2017, pp. 1260−1272.

[187] Q. Yuan, G. Cong, Z. Y. Ma, A. X. Sun, and N. Magnenat-Thalmann, Who, where, when and what: Discover spatio-temporal topics for twitter users, in *Proc. 19th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Chicago, IL, USA, 2013, pp. 605−613.

[188] L. J. Hong, A. Ahmed, S. Gurumurthy, A. J. Smola, and K. Tsioutsiouliklis, Discovering geographical topics in the twitter stream, in *Proc. 21st Int. Conf. on World Wide Web*, Lyon, France, 2012, pp. 769−778.

[189] C. Zhang, K. Y. Zhang, Q. Yuan, L. M. Zhang, T. Hanratty, and J. W. Han, Gmove: Group-level mobility modeling using geo-tagged social media, in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1305−1314.

[190] C. Zhang, L. Y. Liu, D. M. Lei, Q. Yuan, H. L. Zhuang, T. Hanratty, and J. W. Han, Triovecevent: Embedding-based online local event detection in geo-tagged tweet streams, in *Proc. 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Halifax, Canada, 2017, pp. 595−604.

[191] C. Li and D. Goldwasser, Encoding social information with graph convolutional networks forpolitical perspective detection in news media, in *Proc. 57th Ann. Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 2594-2604.

[192] P. Stefanov, K. Darwish, A. Atanasov, and P. Nakov, Predicting the topical stance and political leaning of media using tweets, in *Proc. 58th Ann. Meeting of the Association for Computational Linguistics*, Seattle, WA, USA, 2020, pp. 527−537.

[193] T. H. Nguyen and K. Shirai, Topic modeling based sentiment analysis on social media for stock market prediction, in *Proc. 53rd Ann. Meeting of the Association for Computational Linguistics and the 7th Int. Joint Conf. on Natural Language Processing* (*Volume 1*: *Long Papers*), Beijing, China, 2015, pp. 1354−1364.

[194] B. O'Connor, B. M. Stewart, and N. A. Smith, Learning to extract international relations from political context, in *Proc. 51st Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Sofia, Bulgaria, 2013, pp. 1094−1104.

[195] O. Tsur, D. Calacci, and D. Lazer, A frame of mind: Using statistical models for detection of framing and agenda setting campaigns, in *Proc. 53rd Ann. Meeting of the Association for Computational Linguistics and the 7th Int. Joint Conf. on Natural Language Processing* (*Volume 1*: *Long Papers*), Beijing, China, 2015, pp. 1629−1638.

[196] G. Da San Martino, S. Shaar, Y. F. Zhang, S. Yu, A. Barrón-Cedeño, and P. Nakov, Prta: A system to support the analysis of propaganda techniques in the news, in *Proc. 58th Ann. Meeting of the Association for Computational Linguistics*: *System Demonstrations*, Seattle, WA, USA, 2020, pp. 287−293.

[197] M. Silva, L. S. de Oliveira, A. Andreou, P. O. S. V. de Melo, O. Goga, and F. Benevenuto, Facebook ads monitor: An independent auditing system for political ads on facebook, in *Proc. Web Conf. 2020*, Taipei, China, 2020, pp. 224−234.

[198] B. Cao, L. Zheng, C. W. Zhang, P. S. Yu, A. Piscitello, J. Zulueta, O. Ajilore, K. Ryan, and A. D. Leow, DeepMood: Modeling mobile phone typing dynamics for mood detection, in *Proc. 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Halifax, Canada, 2017, pp. 747−755.

[199] M. Gaur, A. Alambo, J. P. Sain, U. Kursuncu, K. Thirunarayan, R. Kavuluru, A. P. Sheth, R. S. Welton, and J. Pathak, Knowledge-aware assessment of severity of suicide risk for early intervention, in *Proc. World Wide Web Conf.*, San Francisco, CA, USA, 2019, pp. 514−525.

[200] W. H. Yu, M. X. Yu, T. Zhao, and M. Jiang, Identifying referential intention with heterogeneous contexts, in *Proc. Web Conf. 2020*, Taipei, China, 2020, pp. 962−972.

[201] D. Susan, Latent semantic analysis, *Annual Review of Information Science and Technology*, vol. 38, no. 1, pp. 188–230, 2004.

[202] J. Y. Jiang, X. Sun, W. Wang, and S. Young, Enhancing air quality prediction with social media and natural language processing, in *Proc. 57th Ann. Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 2627−2632.

[203] L. Y. Yang, T. L. J. Ng, B. Smyth, and R. Dong, HTML: Hierarchical transformer-based multi-task learning for volatility prediction, in *Proc. Web Conf. 2020*, Taipei, China, 2020, pp. 441-451.

[204] Y. M. Xu and S. B. Cohen, Stock movement prediction from tweets and historical prices, in *Proc. 56th Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Melbourne, Australia, 2018, pp. 1970-1979.

[205] G. M. Liu, T. T. Nguyen, G. Zhao, W. Zha, J. B. Yang, J. N. Cao, M. Wu, P. L. Zhao, and W. Chen, Repeat buyer

prediction for E- commerce, in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 155-164.

[206] S. Chakraborty, A. Venkataraman, S. Jagabathula, and L. Subramanian, Predicting socio-economic indicators using news events, in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1455-1464.

[207] G. Doyle, D. Yurovsky, and M. C. Frank, A robust framework for estimating linguistic alignment in twitter conversations, in *Proc. 25th Int. Conf. on World Wide Web*, Montreal, Canada, 2016, pp. 637-648.

[208] V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena, Statistically significant detection of linguistic change, in *Proc. 24th Int. Conf. on World Wide Web*, Florence, Italy, 2015, pp. 625-635.

[209] H. Gonen, G. Jawahar, D. Seddah, and Y. Goldberg, Simple, interpretable and stable method for detecting words with usage change across corpora, in *Proc. 58th Ann. Meeting of the Association for Computational Linguistics*, Seattle, WA, USA, 2020, pp. 538-555.

[210] J. Tang, S. Wu, B. Gao, and Y. Wan, Topic-level social network search, in *Proc. 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Diego, CA, USA, 2011, pp. 769-772.

[211] D. M. Romero, B. Meeder, and J. M. Kleinberg, Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter, in *Proc. 20th Int. Conf. on World Wide Web*, Hyderabad, India, 2011, pp. 695-704.

[212] B. Meeder, B. Karrer, A. Sayedi, R. Ravi, C. Borgs, and J. T. Chayes, We know who you followed last summer: Inferring social link creation times in twitter, in *Proc. 20th Int. Conf. on World Wide Web*, Hyderabad, India, 2011, pp. 517-526.

[213] S. M. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts, Who says what to whom on twitter, in *Proc. 20th Int. Conf. on World Wide Web*, Hyderabad, India, 2011, pp. 705-714.

[214] K. Lewis, M. Gonzalez, and J. Kaufman, Social selection and peer influence in an online social network, *Proc. Natl. Acad. Sci. USA*, vol. 109, no. 1, pp. 68–72, 2012.

[215] A. L. Schmidt, F. Zollo, M. Del Vicario, A. Bessi, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, Anatomy of news consumption on Facebook, *Proc. Natl. Acad. Sci. USA* , vol. 114, no. 12, pp. 3035–3039, 2017.

[216] D. Eckles, R. F. Kizilcec, and E. Bakshy, Estimating peer effects in networks with peer encouragement designs, *Proc. Natl. Acad. Sci. USA*, vol. 113, no. 27, pp. 7316–7322, 2016.

[217] I. Kayes, N. Kourtellis, D. Quercia, A. Iamnitchi, and F. Bonchi, The social world of content abusers in community question answering, in *Proc. 24th Int. Conf. on World Wide Web*, Florence, Italy, 2015, pp. 570-580.

[218] J. Z. Qiu, Y. X. Li, J. Tang, Z. Lu, H. Ye, B. Chen, Q. Yang, and J. E. Hopcroft, The lifecycle and cascade of wechat social messaging groups, in *Proc. 25th Int. Conf. on World Wide Web*, Montreal, Canada, 2016, pp. 311-320.

[219] K. Lewis, The limits of racial prejudice, *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 47, pp. 18814–18819, 2013.

[220] L. Glowacki, A. Isakov, R. W. Wrangham, R. McDermott, J. H. Fowler, and N. A. Christakis, Formation of raiding parties for intergroup violence is mediated by social network structure, *Proc. Natl. Acad. Sci. USA*, vol. 113, no. 43, pp. 12114–12119, 2016.

[221] D. Braha, Patterns of ties in problem-solving networks and their dynamic properties, *Sci. Rep.* , vol. 10, no. 1, p. 18137, 2020.

[222] R. Dakin and T. B. Ryder, Reciprocity and behavioral heterogeneity govern the stability of social networks, *Proc. Natl. Acad. Sci. USA*, vol. 117, no. 6, pp. 2993–2999, 2020.

[223] M. Gupte, P. Shankar, J. Li, S. Muthukrishnan, and L. Iftode, Finding hierarchy in directed online social networks, in *Proc. 20th Int. Conf. on World Wide Web*, Hyderabad, India, 2011, pp. 557-566.

[224] X. Liu and K. Aberer, Soco: A social network aided context-aware recommender system, in *Proc. 22nd Int. Conf. on World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 781-802.

[225] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, The spreading of misinformation online, *Proc. Natl. Acad. Sci. USA*, vol. 113, no. 3, pp. 554–559, 2016.

[226] J. Lulewicz, The social networks and structural variation of Mississippian sociopolitics in the southeastern United States, *Proc. Natl. Acad. Sci. USA* , vol. 116, no. 14, pp. 6707–6712, 2019.

[227] J. H. Fowler, J. E. Settle, and N. A. Christakis, Correlated genotypes in friendship networks, *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 5, pp. 1993–1997, 2011.

[228] A. I. Roberts and S. G. B. Roberts, Wild chimpanzees modify modality of gestures according to the strength of social bonds and personal network size, *Sci. Rep.*, vol. 6, no. 1, p. 33864, 2016.

[229] K. Hilger, M. Ekman, C. J. Fiebach, and U. Basten, Intelligence is associated with the modular structure of intrinsic brain networks, *Sci. Rep.*, vol. 7, no. 1, p. 16088, 2017.

[230] H. Youn, L. Sutton, E. Smith, C. Moore, J. F. Wilkins, I. Maddieson, W. Croft, and T. Bhattacharya, On the universal structure of human lexical semantics, *Proc. Natl. Acad. Sci. USA*, vol. 113, no. 7, pp. 1766–1771, 2016.

[231] A. E. Sizemore, E. A. Karuza, C. Giusti, and D. S. Bassett, Knowledge gaps in the early growth of semantic feature networks, *Nat. Hum. Behav.* , vol. 2, no. 9, pp. 682–692, 2018.

[232] S. Ronen, B. Gonçalves, K. Z. Hu, A. Vespignani, S. Pinker, and C. A. Hidalgo, Links that speak: The global language network and its association with global fame, *Proc. Natl. Acad. Sci. USA*, vol. 111, no. 52, pp. E5616–E5622, 2014.

[233] L. Taruffi, C. Pehrs, S. Skouras, and S. Koelsch, Effects of sad and happy music on mind-wandering and the default mode network, *Sci. Rep.*, vol. 7, no. 1, p. 14396, 2017.

[234] R. Schmälzle, M. B. O'Donnell, J. O. Garcia, C. N. Cascio, J. Bayer, D. S. Bassett, J. M. Vettel, and E. B. Falk, Brain connectivity dynamics during social

interaction reflect social network structure, *Proc. Natl. Acad. Sci. USA*, vol. 114, no. 20, pp. 5153–5158, 2017.

[235] S. A. Morelli, D. C. Ong, R. Makati, M. O. Jackson, and J. Zaki, Empathy and well-being correlate with centrality in different social networks, *Proc. Natl. Acad. Sci. USA*, vol. 114, no. 37, pp. 9843–9847, 2017.

[236] H. Kim, S. Kwak, J. Kim, Y. Youm, and J. Chey, Social network position moderates the relationship between late-life depressive symptoms and memory differently in men and women, *Sci. Rep.*, vol. 9, no. 1, p. 6142, 2019.

[237] T. Ito, The influence of networks of general trust on willingness to communicate in English for Japanese people, *Sci. Rep.*, vol. 10, no. 1, p. 19939, 2020.

[238] J. Bao, T. F. He, S. J. Ruan, Y. H. Li, and Y. Zheng, Planning bike lanes based on sharing-bikes' trajectories, in *Proc. 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Halifax, Canada, 2017, pp. 1377-1386.

[239] C. Barrington-Leigh and A. Millard-Ball, Global trends toward urban street-network sprawl, *Proc. Natl. Acad. Sci. USA*, vol. 117, no. 4, pp. 1941–1950, 2020.

[240] N. Yadav, S. Chatterjee, and A. R. Ganguly, Resilience of urban transport network-of-networks under intense flood hazards exacerbated by targeted attacks, *Sci. Rep.*, vol. 10, no. 1, p. 10350, 2020.

[241] G. Bonaccorsi, F. Pierri, M. Cinelli, A. Flori, A. Galeazzi, F. Porcelli, A. L. Schmidt, C. M. Valensise, A. Scala, W. Quattrociocchi, et al., Economic and social consequences of human mobility restrictions under COVID-19, *Proc. Natl. Acad. Sci. USA* , vol. 117, no. 27, pp. 15530–15535, 2020.

[242] P. Santi, G. Resta, M. Szell, S. Sobolevsky, S. H. Strogatz, and C. Ratti, Quantifying the benefits of vehicle pooling with shareability networks, *Proc. Natl. Acad. Sci. USA*, vol. 111, no. 37, pp. 13290–13294, 2014.

[243] M. M. Vazifeh, P. Santi, G. Resta, S. H. Strogatz, and C. Ratti, Addressing the minimum fleet problem in on-demand urban mobility, *Nature* , vol. 557, no. 7706, pp. 534–538, 2018.

[244] J. M. Liu, L. L. Sun, W. W. Chen, and H. Xiong, Rebalancing bike sharing systems: A multi-source data smart optimization, in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1005-1014.

[245] A. P. Riascos and J. L. Mateos, Networks and long-range mobility in cities: A study of more than one billion taxi trips in New York City, *Sci. Rep.*, vol. 10, no. 1, p. 4022, 2020.

[246] A. Karduni, A. Kermanshah, and S. Derrible, A protocol to convert spatial polyline data to network formats and applications to world urban road networks, *Sci. Data*, vol. 3, no. 1, p. 160046, 2016.

[247] R. Kujala, C. Weckström, R. K. Darst, M. N, Mladenović, and J. Saramäki, A collection of public transport network data sets for 25 cities, *Sci. Data*, vol. 5, no. 1, p. 180089, 2018.

[248] D. S. Wang, D. Pedreschi, C. M. Song, F. Giannotti, and A. L. Barabási, Human mobility, social ties, and link prediction, in *Proc. 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Diego, CA, USA, 2011, pp. 1100−1108.

[249] E. Cho, S. A. Myers, and J. Leskovec, Friendship and mobility: User movement in location-based social networks, in *Proc. 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Diego, CA, USA, 2011, pp. 1082-1090.

[250] R. Li, S. J. Wang, H. B. Deng, R. Wang, and K. C. C. Chang, Towards social user profiling: Unified and discriminative influence model for inferring home locations, in *Proc. 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Beijing, China, 2012, pp. 1023-1031.

[251] D. N. Yang, C. Y. Shen, W. C. Lee, and M. S. Chen, On socio-spatial group query for location-based social networks, in *Proc. 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Beijing, China, 2012, pp. 949-957.

[252] L. J. Sun, K. W. Axhausen, D. H. Lee, and X. F. Huang, Understanding metropolitan patterns of daily encounters, *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 34, pp. 13774–13779, 2013.

[253] V. Sekara, A. Stopczynski, and S. Lehmann, Fundamental structures of dynamic social networks, *Proc. Natl. Acad. Sci. USA*, vol. 113, no. 36, pp. 9977–9982, 2016.

[254] A. Halu, A. Scala, A. Khiyami, and M. C. González, Data-driven modeling of solar-powered urban microgrids, *Sci. Adv.*, vol. 2, no. 1, p. e1500700, 2016.

[255] S. Battiston, G. Caldarelli, R. M. May, T. Roukny, and J. E. Stiglitz, The price of complexity in financial networks, *Proc. Natl. Acad. Sci. USA*, vol. 113, no. 36, pp. 10031–10036, 2016.

[256] M. Bardoscia, S. Battiston, F. Caccioli, and G. Caldarelli, Pathways towards instability in financial networks, *Nat. Commun.*, vol. 8, no. 1, p. 14416, 2017.

[257] L. L. Porfirio, D. Newth, J. J. Finnigan, and Y. Y. Cai, Economic shifts in agricultural production and trade due to climate change, *Palgrave Commun.* , vol. 4, no. 1, p. 111, 2018.

[258] Z. M. Ren, A. Zeng, and Y. C. Zhang, Bridging nestedness and economic complexity in multilayer world trade networks, *Humanit. Soc. Sci. Commun.* , vol. 7, no. 1, p. 156, 2020.

[259] K. A. Anderson, Skill networks and measures of complex human capital, *Proc. Natl. Acad. Sci. USA* , vol. 114, no. 48, pp. 12720–12724, 2017.

[260] C. A. Bail, L. P. Argyle, T. W. Brown, J. P. Bumpus, H. H. Chen, M. B. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky, Exposure to opposing views on social media can increase political polarization, *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 37, pp. 9216–9221, 2018.

[261] S. Padó, A. Blessing, N. Blokker, E. Dayanik, S. Haunss, and J. Kuhn, Who sides with whom? Towards computational construction of discourse networks for political debates, in *Proc. 57th Ann. Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 2841−2847.

[262] K. Johnson, D. Jin, and D. Goldwasser, Leveraging behavioral and social information for weakly supervised collective classification of political discourse on twitter, in *Proc. 55th Ann. Meeting of the Association for*

Computational Linguistics (*Volume 1*: *Long Papers*), Vancouver, Canada, 2017, pp. 741−752.

[263] A. Bovet and H. A. Makse, Influence of fake news in Twitter during the 2016 US presidential election, *Nat. Commun.*, vol. 10, no. 1, p. 7, 2019.

[264] S. Volkova, G. Coppersmith, and B. Van Durme, Inferring user political preferences from streaming communications, in *Proc. 52nd Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Baltimore, MD, USA, 2014, pp. 186−196.

[265] M. L. Barnes, J. Lynham, K. Kalberg, and P. S. Leung, Social networks and environmental outcomes, *Proc. Natl. Acad. Sci. USA*, vol. 113, no. 23, pp. 6466–6471, 2016.

[266] R. Cámara-Leret, M. A. Fortuna, and J. Bascompte, Indigenous knowledge networks in the face of global change, *Proc. Natl. Acad. Sci. USA* , vol. 116, no. 20, pp. 9913–9918, 2019.

[267] Y. Zheng, F. R. Liu, and H. P. Hsieh, U-air: When urban air quality inference meets big data, in *Proc. 19th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Chicago, IL, USA, 2013, pp. 1436−1444.

[268] H. P. Hsieh, S. D. Lin, and Y. Zheng, Inferring air quality for station location recommendation based on urban big data, in *Proc. 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Sydney, Australia, 2015, pp. 437−446.

[269] A. Guille and H. Hacid, A predictive model for the temporal dynamics of information diffusion in online social networks, in *Proc. 21st Int. Conf. on World Wide Web*, Lyon, France, 2012, pp. 1145-1152.

[270] J. Gómez-Gardeñes, L. Lotero, S. N. Taraskin, and F. J. Pérez-Reche, Explosive contagion in networks, *Sci. Rep.*, vol. 6, no. 1, p. 19767, 2016.

[271] D. Y. Zhang, Y. Wang, and Z. X. Zhang, Identifying and quantifying potential super-spreaders in social networks, *Sci. Rep.*, vol. 9, no. 1, p. 14811, 2019.

[272] L. Gao, C. M. Song, Z. Y. Gao, A. L. Barabási, J. P. Bagrow, and D. S. Wang, Quantifying information flow during emergencies, *Sci. Rep.*, vol. 4, no. 1, p. 3997, 2014.

[273] W. Quattrociocchi, G. Caldarelli, and A. Scala, Opinion dynamics on interacting networks: Media competition and social influence, *Sci. Rep.*, vol. 4, no. 1, p. 4938, 2014.

[274] C. C. Shao, G. L. Ciampaglia, O. Varol, K. C. Yang, A. Flammini, and F. Menczer, The spread of low-credibility content by social bots, *Nat. Commun.* , vol. 9, no. 1, p. 4787, 2018.

[275] S. Vosoughi, D. Roy, and S. Aral, The spread of true and false news online, *Science* , vol. 359, no. 6380, pp. 1146–1151, 2018.

[276] A. Lima, M. De Domenico, V. Pejovic, and M. Musolesi, Disease containment strategies based on mobility and information dissemination, *Sci. Rep.* , vol. 5, no. 1, p. 10650, 2015.

[277] S. J. Luo, F. Morone, C. Sarraute, M. Travizano, and H. A. Makse, Inferring personal economic status from social network location, *Nat. Commun.* , vol. 8, no. 1, p. 15227, 2017.

[278] C. J. Gomez and D. M. J. Lazer, Clustering knowledge and dispersing abilities enhances collective problem solving in a network, *Nat. Commun.* , vol. 10, no. 1, p. 5146, 2019.

[279] X. F. Tang, Y. Z. Liu, N. Shah, X. L. Shi, P. Mitra, and S. H. Wang, Knowing your FATE: Friendship, action and temporal explanations for user engagement prediction on social apps, in *Proc. 26th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, Virtual Event, CA, USA, 2020, pp. 2269−2279.

[280] C. Li, J. Q. Ma, X. X. Guo, and Q. Z. Mei, DeepCas: An end-to-end predictor of information cascades, in *Proc. 26th Int. Conf. on World Wide Web*, Perth, Australia, 2017, pp. 577−586.

[281] J. Z. Qiu, J. Tang, H. Ma, Y. X. Dong, K. S. Wang, and J. Tang, Deepinf: Social influence prediction with deep learning, in *Proc. 24th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, London, UK, 2018, pp. 2110−2119.

[282] Y. J. Lu and C. T. Li, GCAN: Graph-aware co-attention networks for explainable fake news detection on social media, in *Proc. 58th Ann. Meeting of the Association for Computational Linguistics*, Seattle, WA, USA, 2020, pp. 505−514.

[283] L. Zhong, J. Cao, Q. Sheng, J. B. Guo, and Z. Wang, Integrating semantic and structural information with graph convolutional network for controversy detection, in *Proc. 58th Ann. Meeting of the Association for Computational Linguistics*, Seattle, WA, USA, 2020, pp. 515−526.

[284] X. Y. Wang, Y. Ma, Y. Q. Wang, W. Jin, X. Wang, J. L. Tang, C. Y. Jia, and J. Yu, Traffic flow prediction via spatial temporal graph neural network, in *Proc. Web Conf. 2020*, Taipei, China, 2020, pp. 1082−1092.

[285] D. X. Deng, C. Shahabi, U. Demiryurek, L. H. Zhu, R. Yu, and Y. Liu, Latent space model for road networks to predict time-varying traffic, in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1525−1534.

[286] Y. G. Li, K. Fu, Z. Wang, C. Shahabi, J. P. Ye, and Y. Liu, Multi-task representation learning for travel time estimation, in *Proc. 24th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, London, UK, 2018, pp. 1695−1704.

[287] Y. Sun, H. S. Zhu, F. Z. Zhuang, J. J. Gu, and Q. He, Exploring the urban region-of-interest through the analysis of online map search queries, in *Proc. 24th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, London, UK, 2018, pp. 2269−2278.

[288] Z. Y. Pan, Y. X. Liang, W. F. Wang, Y. Yu, Y. Zheng, and J. B. Zhang, Urban traffic prediction from spatio-temporal data using deep meta learning, in *Proc. 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, 2019, pp. 1720−1730.

[289] J. Yuan, Y. Zheng, and X. Xie, Discovering regions of different functions in a city using human mobility and POIs, in *Proc. 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Beijing, China, 2012, pp. 186−194.

[290] P. Y. Wang, Y. J. Fu, J. W. Zhang, P. F. Wang, Y. Zheng, and C. C. Aggarwal, You are how you drive: Peer and temporal-aware representation learning for driving behavior analysis, in *Proc. 24th ACM SIGKDD Int. Conf.*

on Knowledge Discovery & Data Mining, London, UK, 2018, pp. 2457−2466.

[291] Q. W. Zhong, Y. Liu, X. Ao, B. B. Hu, J. H. Feng, J. Y. Tang, and Q. He, Financial defaulter detection on online credit payment via multi-view attributed heterogeneous information network, in *Proc. Web Conf. 2020*, Taipei, China, 2020, pp. 785−795.

[292] J. B. Shang, Y. Zheng, W. Z. Tong, E. Chang, and Y. Yu, Inferring gas consumption and pollution emission of vehicles throughout a city, in *Proc. 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, New York, NY, USA, 2014, pp. 1027−1036.

[293] H. Lakkaraju and J. Ajmera, Attention prediction on social media brand pages, in *Proc. 20th ACM Int. Conf. on Information and Knowledge Management*, Glasgow, UK, 2011, pp. 2157−2160.

[294] S. Volkova and J. Y. Jang, Misleading or falsification: Inferring deceptive strategies and types in online news and social media, in *Proc. Web Conf. 2018*, Lyon, France, 2018, pp. 575−583.

[295] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto, and J. M. Serna, I read but don't agree: Privacy policy benchmarking using machine learning and the EU GDPR, in *Proc. Web Conf. 2018*, Lyon, France, 2018, pp. 163−166.

[296] R. Parimi and D. Caragea, Predicting friendship links in social networks using a topic modeling approach, in *Proc. 15th Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, Shenzhen, China, 2011, pp. 75−86.

[297] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, R. P. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz, *et al*., `Beating the news' with EMBERS: Forecasting civil unrest using open source indicators, in *Proc. 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, New York City, NY, USA, 2014, pp. 1799−1808.

[298] F. Chen and D. B. Neill, Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs, in *Proc. 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, New York, NY, USA, 2014, pp. 1166−1175.

[299] K. Q. Li, W. Lu, S. Bhagat, L. V. S. Lakshmanan, and C. Yu, On social event organization, in *Proc. 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, New York, NY, USA, 2014, pp. 1206−1215.

[300] S. Rayana and L. Akoglu, Collective opinion spam detection: Bridging review networks and metadata, in *Proc. 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Sydney, Australia, 2015, pp. 985−994.

[301] O. Goga, P. Loiseau, R. Sommer, R. Teixeira, and K. P. Gummadi, On the reliability of profile matching across large online social networks, in *Proc. 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Sydney, Australia, 2015, pp. 1799−1808.

[302] M. Madaio, S. T. Chen, O. L. Haimson, W. W. Zhang, X. Cheng, M. Hinds-Aldrich, D. H. Chau, and B Dilkina, Firebird: Predicting fire risk and prioritizing fire inspections in Atlanta, in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 185−194.

[303] B. Shi, W. Lam, L. D. Bing, and Y. Q. Xu, Detecting common discussion topics across culture from news reader comments, in *Proc. 54th Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Berlin, Germany, 2016, pp. 676−685.

[304] S. Bergsma and B. Van Durme, Using conceptual class attributes to characterize social media users, in *Proc. 51st Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Sofia, Bulgaria, 2013, pp. 710−720.

[305] Z. Kozareva, Multilingual affect polarity and valence prediction in metaphor-rich texts, in *Proc. 51st Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Sofia, Bulgaria, 2013, pp. 682−691.

[306] S. Rosenthal and K. McKeown, Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations, in *Proc. 49th Ann. Meeting of the Association for Computational Linguistics*: *Human Language Technologies*, Portland, OR, USA, 2011, pp. 763−772.

[307] S. Park, K. S. Lee, and J. Song, Contrasting opposing views of news articles on contentious issues, in *Proc. 49th Ann. Meeting of the Association for Computational Linguistics*: *Human Language Technologies*, Portland, OR, USA, 2011, pp. 340−349.

[308] C. Castillo, M. Mendoza, and B. Poblete, Information credibility on twitter, in *Proc. 20th Int. Conf. on World Wide Web*, Hyderabad, India, 2011, pp. 675−684.

[309] J. J. Lin, R. Snow, and W. Morgan, Smoothing techniques for adaptive online language models: Topic tracking in tweet streams, in *Proc. 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Diego, CA, USA, 2011, pp. 422−429.

[310] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, A Stylometric inquiry into hyperpartisan and fake news, in *Proc. 56th Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Melbourne, Australia, 2018, pp. 231−240.

[311] J. Ma, W. Gao, and K. F. Wong, Detect rumors in microblog posts using propagation structure via kernel learning, in *Proc. 55th Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Vancouver, Canada, 2017, pp. 708−717.

[312] N. Hassan, F. Arslan, C. K. Li, and M. Tremayne, Toward automated fact- checking: Detecting check-worthy factual claims by ClaimBuster, in *Proc. 23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Halifax, Canada, 2017, pp. 1803−1812.

[313] Y. X. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla, Inferring user demographics and social strategies in mobile social networks, in *Proc. 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, New York, NY, USA, 2014, pp. 15−24.

[314] B. Hooi, H. A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos, Fraudar: Bounding graph fraud in the face of camouflage, in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 895−904.

[315] C. Budak, D. Agrawal, and A. El Abbadi, Limiting the spread of misinformation in social networks, in *Proc. 20th*

Int. Conf. on World Wide Web, Hyderabad, India, 2011, pp. 665-674.

[316] J. Ratkiewicz, M. D. Conover, M. R. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer, Truthy: Mapping the spread of astroturf in microblog streams, in *Proc. 20th Int. Conf. Companion on World Wide Web*, Hyderabad, India, 2011, pp. 249−252.

[317] M. Jamali, G. Haffari, and M. Ester, Modeling the temporal dynamics of social rating networks using bidirectional effects of social relations and rating patterns, in *Proc. 20th Int. Conf. on World Wide Web*, Hyderabad, India, 2011, pp. 527−536.

[318] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and P. K. Gummadi, Understanding and combating link farming in the twitter social network, in *Proc. 21st Int. Conf. on World Wide Web*, Lyon, France, 2012, pp. 61−70.

[319] C. Yang, R. C. Harkreader, J. L. Zhang, S. Shin, and G. F. Gu, Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on twitter, in *Proc. 21st Int. Conf. on World Wide Web*, Lyon, France, 2012, pp. 71−80.

[320] G. Ver Steeg and A. Galstyan, Information transfer in social media, in *Proc. 21st Int. Conf. on World Wide Web*, Lyon, France, 2012, pp. 509−518.

[321] A. Beutel, W. H. Xu, V. Guruswami, C. Palow, and C. Faloutsos, CopyCatch: Stopping group attacks by spotting lockstep behavior in social networks, in *Proc. 22nd Int. Conf. on World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 119−130.

[322] T. C. Lou and J. Tang, Mining structural hole spanners through information diffusion in social networks, in *Proc. 22nd Int. Conf. on World Wide Web*, Rio de Janeiro, Brazil, 2013, pp. 825−836.

[323] J. Cheng, L. A. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, Can cascades be predicted? in *Proc. 23rd Int. Conf. on World Wide Web*, Seoul, Republic of Korea, 2014, pp. 925−936.

[324] S. A. Myers, A. Sharma, P. Gupta, and J. J. Lin, Information network or social network?: The structure of the Twitter follow graph, in *Proc. 23rd Int. Conf. on World Wide Web*, Seoul, Republic of Korea, 2014, pp. 493−498.

[325] C. Buntain and J. Golbeck, Identifying social roles in Reddit using network structure, in *Proc. 23rd Int. Conf. on World Wide Web*, Seoul, Republic of Korea, 2014, pp. 615-620.

[326] U. Pavalanathan and M. De Choudhury, Identity management and mental health discourse in social media, in *Proc. 24th Int. Conf. on World Wide Web*, Florence, Italy, 2015, pp. 315−321.

[327] P. Singer, D. Helic, A. Hotho, and M. Strohmaier, HypTrails: A bayesian approach for comparing hypotheses about human trails on the web, in *Proc. 24th Int. Conf. on World Wide Web*, Florence, Italy, 2015, pp. 1003−1013.

[328] M. Yin, M. L. Gray, S. Suri, and J. W. Vaughan, The communication network within the crowd, in *Proc. 25th Int. Conf. on World Wide Web*, Montreal, Canada, 2016, pp. 1293−1303.

[329] J. Su, A. Sharma, and S. Goel, The effect of recommendations on network structure, in *Proc. 25th Int. Conf. on World Wide Web*, Montreal, Canada, 2016, pp. 1157−1167.

[330] D. M. Romero, B. Uzzi, and J. Kleinberg, Social networks under stress, in *Proc. 25th Int. Conf. on World Wide Web*, Montreal, Canada, 2016, pp. 9−20.

[331] Á. García-Recuero, Discouraging abusive behavior in privacy-preserving online social networking applications, in *Proc. 25th Int. Conf. Companion on World Wide Web*, Montreal, Canada, 2016, pp. 305−309.

[332] Y. X. Li, O. Martinez, X. Chen, Y. Li, and J. E. Hopcroft, In a world that counts: Clustering and detecting fake social engagement at scale, in *Proc. 25th Int. Conf. on World Wide Web*, Montreal, Canada, 2016, pp. 111−120.

[333] G. Resende, P. F. Melo, H. Sousa, J. Messias, M. Vasconcelos, J. M. Almeida, and F. Benevenuto, (Mis) information dissemination in WhatsApp. Gathering, analyzing and countermeasures, in *Proc. World Wide Web Conf.*, San Francisco, CA, USA, 2019, pp. 818−828.

[334] N. E. Friedkin, A. V. Proskurnikov, R. Tempo, and S. E. Parsegov, Network science on belief system dynamics under logic constraints, *Science* , vol. 354, no. 6310, pp. 321–326, 2016.

[335] S. Mukherjee, D. M. Romero, B. Jones, and B. Uzzi, The nearly universal link between the age of past knowledge and tomorrow's breakthroughs in science and technology: The hotspot, *Sci. Adv.*, vol. 3, no. 4, p. e1601315, 2017.

[336] X. Jin, C. Wang, J. B. Luo, X. Yu, and J. W. Han, LikeMiner: A system for mining the power of 'like' in social media networks, in *Proc. 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Diego, CA, USA, 2011, pp. 753−756.

[337] P. Papadimitriou, H. Garcia-Molina, P. Krishnamurthy, R. A. Lewis, and D. H. Reiley, Display advertising impact: Search lift and social influence, in *Proc. 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Diego, CA, USA, 2011, pp. 1019−1027.

[338] C. T. Li and S. D. Lin, Social flocks: A crowd simulation framework for social network generation, community detection, and collective behavior modeling, in *Proc. 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Diego, CA, USA, 2011, pp. 765−768.

[339] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos, Rise and fall patterns of information diffusion: Model and implications, in *Proc. 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Beijing, China, 2012, pp. 6−14.

[340] D. G. Rand, S. Arbesman, and N. A. Christakis, Dynamic social networks promote cooperation in experiments with humans, *Proc. Natl. Acad. Sci. USA* , vol. 108, no. 48, pp. 19193–19198, 2011.

[341] G. Facchetti, G. Iacono, and C. Altafini, Computing global structural balance in large-scale signed social networks, *Proc. Natl. Acad. Sci. USA* , vol. 108, no. 52, pp. 20953–20958, 2011.

[342] C. P. Roca and D. Helbing, Emergence of social cohesion in a model society of greedy, mobile individuals, *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 28, pp. 11370–11374, 2011.

[343] L. Dall'Asta, M. Marsili, and P. Pin, Collaboration in

social networks, *Proc. Natl. Acad. Sci. USA* , vol. 109, no. 12, pp. 4395–4400, 2012.

[344] B. J. Mills, J. J. Clark, M. A. Peeples, W. R. Haas Jr, J. M. Roberts Jr, J. B. Hill, D. L. Huntley, L. Borck, R. L. Breiger, A. Clauset, et al., Transformation of social networks in the late pre-Hispanic US Southwest, *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 15, pp. 5785–5790, 2013.

[345] Z. Q. Jiang, W. J. Xie, M. X. Li, B. Podobnik, W. X. Zhou, and H. E. Stanley, Calling patterns in human communication dynamics, *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 5, pp. 1600–1605, 2013.

[346] A. Rutherford, M. Cebrian, S. Dsouza, E. Moro, A. Pentland, and I. Rahwan, Limits of social mobilization, *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 16, pp. 6281–6286, 2013.

[347] J. Saramäki, E. A. Leicht, E. López, S. G. B. Roberts, F. Reed-Tsochas, and R. I. M. Dunbar, Persistence of social signatures in human communication, *Proc. Natl. Acad. Sci. USA*, vol. 111, no. 3, pp. 942–947, 2014.

[348] A. Rzhetsky, J. G. Foster, I. T. Foster, and J. A. Evans, Choosing experiments to accelerate collective discovery, *Proc. Natl. Acad. Sci. USA*, vol. 112, no. 47, pp. 14569–14574, 2015.

[349] A. M. Petersen, Quantifying the impact of weak, strong, and super ties in scientific careers, *Proc. Natl. Acad. Sci. USA*, vol. 112, no. 34, pp. E4671–E4680, 2015.

[350] E. L. Paluck, H. Shepherd, and P. M. Aronow, Changing climates of conflict: A social network experiment in 56 schools, *Proc. Natl. Acad. Sci. USA* , vol. 113, no. 3, pp. 566–571, 2016.

[351] A. Coman, I. Momennejad, R. D. Drach, and A. Geana, Mnemonic convergence in social networks: The emergent properties of cognition at a collective level, *Proc. Natl. Acad. Sci. USA*, vol. 113, no. 29, pp. 8171–8176, 2016.

[352] X. Han, S. N. Cao, Z. S. Shen, B. Y. Zhang, W. X. Wang, R. Cressman, and H. E. Stanley, Emergence of communities and diversity in social networks, *Proc. Natl. Acad. Sci. USA*, vol. 114, no. 11, pp. 2887–2891, 2017.

[353] D. Guilbeault, J. Becker, and D. Centola, Social learning and partisan bias in the interpretation of climate trends, *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 39, pp. 9714–9719, 2018.

[354] I. Tamarit, J. A. Cuesta, R. I. M. Dunbar, and A. Sánchez, Cognitive resource allocation determines the organization of personal networks, *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 33, pp. 8316–8321, 2018.

[355] C. Stadtfeld, A. Vörös, T. Elmer, Z. Boda, and I. J. Raabe, Integration in emerging social networks explains academic failure and success, *Proc. Natl. Acad. Sci. USA*, vol. 116, no. 3, pp. 792–797, 2019.

[356] Y. Yang, N. V. Chawla, and B. Uzzi, A network's gender composition and communication pattern predict women's leadership success, *Proc. Natl. Acad. Sci. USA* , vol. 116, no. 6, pp. 2033–2038, 2019.

[357] D. R. Lo Sardo, S. Thurner, J. Sorger, G. Duftschmid, G. Endel, and P. Klimek, Quantification of the resilience of primary care networks by stress testing the health care system, *Proc. Natl. Acad. Sci. USA* , vol. 116, no. 48, pp. 23930–23935, 2019.

[358] A. Almaatouq, A. Noriega-Campero, A. Alotaibi, P. M. Krafft, M. Moussaid, and A. Pentland, Adaptive social networks promote the wisdom of crowds, *Proc. Natl. Acad. Sci. USA*, vol. 117, no. 21, pp. 11379–11386, 2020.

[359] N. Rovira-Asenjo, T. Gumí, M. Sales-Pardo, and R. Guimerà, Predicting future conflict between team-members with parameter-free models of social networks, *Sci. Rep.*, vol. 3, no. 1, p. 1999, 2013.

[360] M. H. Li, H. L. Zou, S. G. Guan, X. F. Gong, K. Li, Z. R. Di, and C. H. Lai, A coevolving model based on preferential triadic closure for social media networks, *Sci. Rep.*, vol. 3, no. 1, p. 2512, 2013.

[361] W. Wang, Q. H. Liu, S. M. Cai, M. Tang, L. A. Braunstein, and H. E. Stanley, Suppressing disease spreading by using information diffusion on multiplex networks, *Sci. Rep.*, vol. 6, no. 1, p. 29259, 2016.

[362] S. Aral and C. Nicolaides, Exercise contagion in a global social network, *Nat. Commun.* , vol. 8, no. 1, p. 14753, 2017.

[363] M. Del Vicario, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, Modeling confirmation bias and polarization, *Sci. Rep.*, vol. 7, no. 1, p. 40391, 2017.

[364] C. Parkinson, A. M. Kleinbaum, and T. Wheatley, Spontaneous neural encoding of social network position, *Nat. Hum. Behav.*, vol. 1, no. 5, p. 0072, 2017.

[365] F. Battiston, V. Nicosia, V. Latora, and M. S. Miguel, Layered social influence promotes multiculturality in the Axelrod model, *Sci. Rep.*, vol. 7, no. 1, p. 1809, 2017.

[366] C. Shen, C. Chu, H. Guo, L. Shi, and J. Y. Duan, Coevolution of vertex weights resolves social dilemma in spatial networks, *Sci. Rep.*, vol. 7, no. 1, p. 15213, 2017.

[367] Y. E. Wu, S. H. Chang, Z. P. Zhang, and Z. H. Deng, Impact of social reward on the evolution of the cooperation behavior in complex networks, *Sci. Rep.*, vol. 7, no. 1, p. 41076, 2017.

[368] K. M. Altenburger and J. Ugander, Monophily in social networks introduces similarity among friends-of-friends, *Nat. Hum. Behav.*, vol. 2, no. 4, pp. 284–290, 2018.

[369] I. Iacopini, G. Petri, A. Barrat, and V. Latora, Simplicial models of social contagion, *Nat. Commun.*, vol. 10, no. 1, p. 2485, 2019.

[370] N. F. Johnson, R. Leahy, N. J. Restrepo, N. Velasquez, M. Zheng, P. Manrique, P. Devkota, and S. Wuchty, Hidden resilience and adaptive dynamics of the global online hate ecology, *Nature* , vol. 573, no. 7773, pp. 261–265, 2019.

[371] E. Lee, F. Karimi, C. Wagner, H. H. Jo, M. Strohmaier, and M. Galesic, Homophily and minority-group size explain perception biases in social networks, *Nat. Hum. Behav.*, vol. 3, no. 10, pp. 1078–1087, 2019.

[372] R. Schuchard, A. Crooks, A. Stefanidis, and A. Croitoru, Bots fired: Examining social bot evidence in online mass shooting conversations, *Palgrave Commun.*, vol. 5, no. 1, p. 158, 2019.

[373] Z. Q. Zhu, C. Gao, Y. M. Zhang, H. N. Li, J. Xu, Y. L. Zan, and Z. Li, Cooperation and competition among information on social networks, *Sci. Rep.*, vol. 10, no. 1, p. 12160, 2020.

[374] T. David-Barrett, Herding friends in similarity-based architecture of social networks, *Sci. Rep.*, vol. 10, no. 1, p. 4859, 2020.

[375] C. Pomeroy, R. M. Bond, P. J. Mucha, and S. J. Cranmer,

Dynamics of social network emergence explain network evolution, *Sci. Rep.*, vol. 10, no. 1, p. 21876, 2020.
[376] G. Pickard, W. Pan, I. Rahwan, M. Cebrian, R. Crane, A. Madan, and A. Pentland, Time-critical social mobilization, *Science* , vol. 334, no. 6055, pp. 509–512, 2011.
[377] H. P. Young, The dynamics of social innovation, *Proc. Natl. Acad. Sci. USA*, vol. 108, no. S4, pp. 21285–21291, 2011.
[378] N. S. Contractor and L. A. DeChurch, Integrating social networks and human social motives to achieve social influence at scale, *Proc. Natl. Acad. Sci. USA*, vol. 111, no. S4, pp. 13650–13657, 2014.
[379] J. Becker, D. Brackbill, and D. Centola, Network dynamics of social influence in the wisdom of crowds, *Proc. Natl. Acad. Sci. USA*, vol. 114, no. 26, pp. E5070–E5076, 2017.
[380] J. S. Sayles and J. A. Baggio, Social-ecological network analysis of scale mismatches in estuary watershed restoration, *Proc. Natl. Acad. Sci. USA* , vol. 114, no. 10, pp. E1776–E1785, 2017.
[381] Y. Q. Hu, S. G. Ji, Y. L. Jin, L. Feng, H. E. Stanley, and S. Havlin, Local structure can identify and quantify influential global spreaders in large scale social networks, *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 29, pp. 7468–7472, 2018.
[382] Y. F. Ma and B. Uzzi, Scientific prize network predicts who pushes the boundaries of science, *Proc. Natl. Acad. Sci. USA*, vol. 115, no. 50, pp. 12608–12615, 2018.
[383] T. Wei, M. H. Li, C. S. Wu, X. Y. Yan, Y. Fan, Z. R. Di, and J. S. Wu, Do scientists trace hot topics? *Sci. Rep.*, vol. 3, no. 1, p. 2207, 2013.
[384] A. Ehlert, M. Kindschi, R. Algesheimer, and H. Rauhut, Human social preferences cluster and spread in the field, *Proc. Natl. Acad. Sci. USA*, vol. 117, no. 37, pp. 22787–22792, 2020.
[385] F. Gargiulo and T. Carletti, Driving forces of researchers mobility, *Sci. Rep.*, vol. 4, no. 1, p. 4860, 2014.
[386] P. H. C. Guerra, A. Veloso, W. Meira Jr, and V. A. F. Almeida, From bias to opinion: A transfer-learning approach to real-time sentiment analysis, in *Proc. 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Diego, CA, USA, 2011, pp. 150-158.
[387] Y. Y. Wu, M. Kosinski, and D. Stillwell, Computer-based personality judgments are more accurate than those made by humans, *Proc. Natl. Acad. Sci. USA* , vol. 112, no. 4, pp. 1036–1040, 2015.
[388] F. Schlosser, B. F. Maier, O. Jack, D. Hinrichs, A. Zachariae, and D. Brockmann, COVID-19 lockdown induces disease-mitigating structural changes in mobility networks, *Proc. Natl. Acad. Sci. USA* , vol. 117, no. 52, pp. 32883–32890, 2020.
[389] A. D. Henry, P. Prałat, and C. Q. Zhang, Emergence of segregation in evolving social networks, *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 21, pp. 8605–8610, 2011.
[390] W. Mason and D. J. Watts, Collaborative learning in networks, *Proc. Natl. Acad. Sci. USA* , vol. 109, no. 3, pp. 764–769, 2012.
[391] J. Wang, S. Suri, and D. J. Watts, Cooperation and assortativity with dynamic partner updating, *Proc. Natl. Acad. Sci. USA*, vol. 109, no. 36, pp. 14363–14368, 2012.
[392] P. Dandekar, A. Goel, and D. T. Lee, Biased assimilation, homophily, and the dynamics of polarization, *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 15, pp. 5791–5796, 2013.
[393] A. Varga, Shorter distances between papers over time are due to more cross-field references and increased citation rate to higher-impact papers, *Proc. Natl. Acad. Sci. USA*, vol. 116, no. 44, pp. 22094–22099, 2019.
[394] J. Becker, E. Porter, and D. Centola, The wisdom of partisan crowds, *Proc. Natl. Acad. Sci. USA* , vol. 116, no. 22, pp. 10717–10722, 2019.
[395] L. Lucchini, L. Alessandretti, B. Lepri, A. Gallo, and A. Baronchelli, From code to market: Network of developers and correlated returns of cryptocurrencies, *Sci. Adv.*, vol. 6, no. 51, p. eabd2204, 2020.
[396] H. Shirado and N. A. Christakis, Locally noisy autonomous agents improve global human coordination in network experiments, *Nature* , vol. 545, no. 7654, pp. 370–374, 2017.
[397] A. J. Stewart, M. Mosleh, M. Diakonova, A. A. Arechar, D. G. Rand, and J. B. Plotkin, Information gerrymandering and undemocratic decisions, *Nature*, vol. 573, no. 7772, pp. 117–121, 2019.
[398] L. Muchnik, S. Pei, L. C. Parra, S. D. S. Reis, J. S. Andrade Jr, S. Havlin, and H. A. Makse, Origins of power-law degree distribution in the heterogeneity of human activity in social networks, *Sci. Rep.*, vol. 3, no. 1, p. 1783, 2013.
[399] Z. Wang, C. Y. Xia, S. Meloni, C. S. Zhou, and Y. Moreno, Impact of social punishment on cooperative behavior in complex networks, *Sci. Rep.* , vol. 3, no. 1, p. 3055, 2013.
[400] S. F. Lu, G. Z. Jin, B. Uzzi, and B. Jones, The retraction penalty: Evidence from the web of science, *Sci. Rep.*, vol. 3, no. 1, p. 3146, 2013.
[401] P. Singh, S. Sreenivasan, B. K. Szymanski, and G. Korniss, Threshold-limited spreading in social networks with multiple initiators, *Sci. Rep.* , vol. 3, no. 1, p. 2330, 2013.
[402] R. W. Wilkins, D. A. Hodges, P. J. Laurienti, M. Steen, and J. H. Burdette, Network science and the effects of music preference on functional brain connectivity: From Beethoven to Eminem, *Sci. Rep.* , vol. 4, no. 1, p. 6130, 2014.
[403] D. A. Gianetto and B. Heydari, Network modularity is essential for evolution of cooperation under uncertainty, *Sci. Rep.*, vol. 5, no. 1, p. 9340, 2015.
[404] J. A. Cuesta, C. Gracia-Lázaro, A. Ferrer, Y. Moreno, and A. Sánchez, Reputation drives cooperative behaviour and network formation in human groups, *Sci. Rep.*, vol. 5, no. 1, p. 7843, 2015.
[405] M. Ramos, J. Shao, S. D. S. Reis, C. Anteneodo, J. S. Andrade, S. Havlin, and H. A. Makse, How does public opinion become extreme? *Sci. Rep.* , vol. 5, no. 1, p. 10032, 2015.
[406] G. L. Yang, T. P. Benko, M. Cavaliere, J. C. Huang, and M. Perc, Identification of influential invaders in evolutionary populations, *Sci. Rep.*, vol. 9, no. 1, p. 7305, 2019.
[407] R. Zafarani and H. Liu, Connecting users across social media sites: A behavioral-modeling approach, in *Proc. 19th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Chicago, IL, USA, 2013, pp. 41−49.
[408] A. Mukherjee, A. Kumar, B. Liu, J. H. Wang, M. Hsu,

M. Castellanos, and R. Ghosh, Spotting opinion spammers using behavioral footprints, in *Proc. 19th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Chicago, IL, USA, 2013, pp. 632−640.

[409] P. Lucey, D. Oliver, P. Carr, J. Roth, and I. A. Matthews, Assessing team strategy using spatiotemporal data, in *Proc. 19th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Chicago, IL, USA, 2013, pp. 1366−1374.

[410] A. F. Costa, Y. Yamaguchi, A. J. M. Traina, C. Traina Jr, and C. Faloutsos, RSC: Mining and modeling temporal activity in social media, in *Proc. 21th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Sydney, Australia, 2015, pp. 269−278.

[411] X. Mu, F. D. Zhu, E. P. Lim, J. Xiao, J. Z. Wang, and Z. H. Zhou, User identity linkage by latent user space modelling, in *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1775−1784.

[412] Y. Q. Wang, F. L. Ma, Z. W. Jin, Y. Yuan, G. X. Xun, K. Jha, L. Su, and J. Gao, EANN: Event adversarial neural networks for multi-modal fake news detection, in *Proc. 24th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, London, UK, 2018, pp. 849−857.

[413] R. H. Jiang, X. Song, D. Huang, X. Y. Song, T. Q. Xia, Z. K. Cai, Z. N. Wang, K. S. Kim, and R. Shibasaki, DeepUrbanEvent: A system for predicting citywide crowd dynamics at big events, in *Proc. 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, 2019, 2114−2122.

[414] D. Z. Ding, M. Zhang, X. D. Pan, M. Yang, and X. N. He, Modeling extreme events in time series prediction, in *Proc. 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, 2019, pp. 1114−1122.

[415] F. Zarrinkalam, H. Fani, and E. Bagheri, Social user interest mining: Methods and applications, in *Proc. 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, 2019, pp. 3235−3236.

[416] J. Q. Zhang, B. Bai, Y. Lin, J. Liang, K. Bai, and F. Wang, General-purpose user Embeddings based on mobile app usage, in *Proc. 26th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, Virtual Event, CA, USA, 2020, pp. 2831−2840.

[417] S. Dutta, S. Masud, S. Chakrabarti, and T. Chakraborty, Deep exogenous and endogenous influence combination for social chatter intensity prediction, in *Proc. 26th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, Virtual Event, CA, USA, 2020, pp. 1999−2008.

[418] J. T. Ye and S. Skiena, The secret lives of names?: Name embeddings from social media, in *Proc. 25th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, 2019, pp. 3000−3008.

[419] R. Baly, G. Karadzhov, J. S. An, H. Kwak, Y. Dinkov, A. Ali, J. R. Glass, and P. Nakov, What was written vs. who read it: News media profiling using text analysis and social media context, in *Proc. 58th Ann. Meeting of the Association for Computational Linguistics*, Seattle, WA, USA, 2020, pp. 3364−3374.

[420] L. W. Wu, Y. Rao, Y. Q. Zhao, H. Liang, and A. Nazir, DTCA: Decision tree-based co-attention networks for explainable claim verification, in *Proc. 58th Ann. Meeting of the Association for Computational Linguistics*, Seattle, WA, USA, 2020, pp. 1024−1035.

[421] S. Bansal, V. Garimella, A. Suhane, J. Patro, and A. Mukherjee, Code-switching patterns can be an effective route to improve performance of downstream NLP applications: A case study of Humour, sarcasm and hate speech detection, in *Proc. 58th Ann. Meeting of the Association for Computational Linguistics*, Seattle, WA, USA, 2020, pp. 1018−1023.

[422] A. G. Chowdhury, R. Sawhney, R. R. Shah, and D. Mahata, YouToo? Detection of personal recollections of sexual harassment on social media, in *Proc. 57th Ann. Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 2527−2537.

[423] S. Oprea and W. Magdy, Exploring author context for detecting intended vs. perceived sarcasm, in *Proc. 57th Ann. Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 2854−2859.

[424] J. Ma, W. Gao, S. Joty, and K. F. Wong, Sentence-level evidence embedding for claim verification with hierarchical attention networks, in *Proc. 57th Ann. Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 2561−2571.

[425] M. T. Wan, R. Misra, N. Nakashole, and J. McAuley, Fine-grained spoiler detection from large-scale review corpora, in *Proc. 57th Ann. Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019, pp. 2605−2610.

[426] J. Ma, W. Gao, and K. F. Wong, Rumor detection on twitter with tree-structured recursive neural networks, in *Proc. 56th Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Melbourne, Australia, 2018, pp. 1980−1989.

[427] A. Mishra, K. Dey, and P. Bhattacharyya, Learning cognitive features from gaze data for sentiment and sarcasm classification using convolutional neural network, in *Proc. 55th Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Vancouver, Canada, 2017, pp. 377−387.

[428] U. Pavalanathan, J. Fitzpatrick, S. Kiesling, and J. Eisenstein, A multidimensional lexicon for interpersonal stancetaking, in *Proc. 55th Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Vancouver, Canada, 2017, pp. 884−895.

[429] A. Sasaki, K. Hanawa, N. Okazaki, and K. Inui, Other topics you may also agree or disagree: Modeling inter-topic preferences using tweets and matrix factorization, in *Proc. 55th Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Vancouver, Canada, 2017, pp. 398−408.

[430] S. Volkova and Y. Bachrach, Inferring perceived demographics from user emotional tone and user-environment emotional contrast, in *Proc. 54th Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Berlin, Germany, 2016, pp. 1567−1578.

[431] D. Preoţiuc-Pietro, V. Lampos, and N. Aletras, An analysis of the user occupational class through Twitter content, in *Proc. 53rd Ann. Meeting of the Association for*

Computational Linguistics and the 7th Int. Joint Conf. on Natural Language Processing (*Volume 1*: *Long Papers*), Beijing, China, 2015, pp. 1754−1764.

[432] J. W. Li, M. Ott, C. Cardie, and E. Hovy, Towards a general rule for identifying deceptive opinion spam, in *Proc. 52nd Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Baltimore, MD, USA, 2014, pp. 1566−1576.

[433] M. Yancheva and F. Rudzicz, Automatic detection of deception in child-produced speech using syntactic complexity features, in *Proc. 51st Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Sofia, Bulgaria, 2013, pp. 944−953.

[434] Q. M. Diao, J. Jiang, F. D. Zhu, and E. P. Lim, Finding bursty topics from microblogs, in *Proc. 50th Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Jeju Island, Republic of Korea, 2012, pp. 536−544.

[435] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati, Hate speech detection with comment embeddings, in *Proc. 24th Int. Conf. on World Wide Web*, Florence, Italy, 2015, pp. 29−30.

[436] K. H. Lim, K. E. Lee, D. Kendal, L. Rashidi, E. Naghizade, S. Winter, and M. Vasardani, The grass is greener on the other side: Understanding the effects of green spaces on Twitter user sentiments, in *Proc. Web Conf. 2018*, Lyon, France, 2018, pp. 275−282.

[437] K. Johnson and D. Goldwasser, Classification of moral foundations in Microblog political discourse, in *Proc. 56th Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Melbourne, Australia, 2018, pp. 720−730.

[438] V. Lampos, D. Preoţiuc-Pietro, and T. Cohn, A user-centric model of voting intention from social media, in *Proc. 51st Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Sofia, Bulgaria, 2013, pp. 993−1003.

[439] A. Badawy, K. Lerman, and E. Ferrara, Who falls for online political manipulation? in *Proc. 2019 World Wide Web Conf.*, San Francisco, CA, USA, 2019, pp. 162−168.

[440] V. A. Nguyen, J. Boyd-Graber, P. Resnik, and K. Miler, Tea party in the house: A hierarchical ideal point topic model and its application to republican legislators in the 112th congress, in *Proc. 53rd Ann. Meeting of the Association for Computational Linguistics and the 7th Int. Joint Conf. on Natural Language Processing* (*Volume 1*: *Long Papers*), Beijing, China, 2015, pp. 1438−1448.

[441] M. Iyyer, P. Enns, J. Boyd-Graber, and P. Resnik, Political ideology detection using recursive neural networks, in *Proc. 52nd Ann. Meeting of the Association for Computational Linguistics* (*Volume 1*: *Long Papers*), Baltimore, MD, USA, 2014, pp. 1113−1122.

[442] E. Atalay, A. Hortaçsu, J. Roberts, and C. Syverson, Network structure of production, *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 13, pp. 5199–5202, 2011.

[443] N. Arinaminpathy, S. Kapadia, and R. M. May, Size and complexity in model financial systems, *Proc. Natl. Acad. Sci. USA*, vol. 109, no. 45, pp. 18338–18343, 2012.

[444] A. G. Haldane and R. M. May, Systemic risk in banking ecosystems, *Nature* , vol. 469, no. 7330, pp. 351–355, 2011.

[445] F. Pozzi, T. Di Matteo, and T. Aste, Spread of risk across financial markets: Better to invest in the peripheries, *Sci. Rep.*, vol. 3, no. 1, p. 1665, 2013.

[446] T. Squartini, I. Van Lelyveld, and D. Garlaschelli, Early-warning signals of topological collapse in interbank networks, *Sci. Rep.*, vol. 3, no. 1, p. 3357, 2013.

[447] S. Thurner and S. Poledna, DebtRank-transparency: Controlling systemic risk in financial networks, *Sci. Rep.*, vol. 3, no. 1, p. 1888, 2013.

[448] G. Cimini, T. Squartini, D. Garlaschelli, and A. Gabrielli, Systemic risk analysis on reconstructed economic and financial networks, *Sci. Rep.* , vol. 5, no. 1, p. 15758, 2015.

[449] G. L. Ciampaglia, A. Flammini, and F. Menczer, The production of information in the attention economy, *Sci. Rep.*, vol. 5, no. 1, p. 9452, 2015.

[450] O. Filip, K. Janda, L. Kristoufek, and D. Zilberman, Dynamics and evolution of the role of biofuels in global commodity and financial markets, *Nat. Energy* , vol. 1, no. 12, p. 16169, 2016.

[451] J. García-Algarra, M. L. Mouronte-López, and J. Galeano, A stochastic generative model of the world trade network, *Sci. Rep.*, vol. 9, no. 1, p. 18539, 2019.

[452] J. T. Kao, J. Y. Wu, L. Bergen, and N. D. Goodman, Nonliteral understanding of number words, *Proc. Natl. Acad. Sci. USA*, vol. 111, no. 33, pp. 12002–12007, 2014.

[453] A. Bouchard-Côté, D. Hall, T. L. Griffiths, and D. Klein, Automated reconstruction of ancient languages using probabilistic models of sound change, *Proc. Natl. Acad. Sci. USA*, vol. 110, no. 11, pp. 4224–4229, 2013.

[454] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E. Gaughan, V. D. Blondel, and A. J. Tatem, Dynamic population mapping using mobile phone data, *Proc. Natl. Acad. Sci. USA*, vol. 111, no. 45, pp. 15888–15893, 2014.

[455] L. M. A. Bettencourt, The origins of scaling in cities, *Science*, vol. 340, no. 6139, pp. 1438–1441, 2013.

[456] M. Barthelemy, P. Bordin, H. Berestycki, and M. Gribaudi, Self-organization versus top-down planning in the evolution of a city, *Sci. Rep.* , vol. 3, no. 1, p. 2153, 2013.

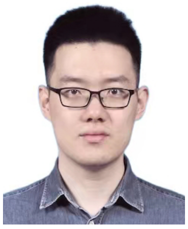
[457] D. Q. Li, Y. N. Jiang, R. Kang, and S. Havlin, Spatial correlation analysis of cascading failures: Congestions and blackouts, *Sci. Rep.*, vol. 4, no. 1, p. 5381, 2014.

[458] Z. Su, L. X. Li, H. P. Peng, J. Kurths, J. H. Xiao, and Y. X. Yang, Robustness of interrelated traffic networks to cascading failures, *Sci. Rep.*, vol. 4, no. 1, p. 5413, 2014.

[459] Y. L. Wang, Y. Zheng, and Y. X, Xue, Travel time estimation of a path using sparse trajectories, in *Proc. 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, New York City, NY, USA, 2014, pp. 25−34.

[460] R. I. McDonald, P. Green, D. Balk, B. M. Fekete, C. Revenga, M. Todd, and M. Montgomery, Urban growth, climate change, and freshwater availability, *Proc. Natl. Acad. Sci. USA*, vol. 108, no. 15, pp. 6312–6317, 2011.

[461] C. Dalin, M. Konar, N. Hanasaki, A. Rinaldo, and I. Rodriguez-Iturbe, Evolution of the global virtual water trade network, *Proc. Natl. Acad. Sci. USA* , vol. 109, no. 16, pp. 5989–5994, 2012.



Cheng Yang received the BEng and PhD degrees from Tsinghua University in 2014 and 2019, respectively. He is currently an assistant professor at Beijing University of Posts and Telecommunications. His research interests include natural language processing and network representation learning. He has published more than 20 top-level papers in international journals and conferences, including ACM TOIS, EMNLP, IJCAI, and AAAI.

Zhiyuan Liu received the BEng and PhD degrees from Tsinghua University in 2006 and 2011, respectively. He is currently an associate professor at the Department of Computer Science and Technology, Tsinghua University. His research interests include natural language processing and social computation. He has published over 80 papers in international journals and conferences, including *ACM Transactions*, IJCAI, AAAI, ACL, and EMNLP.

Maosong Sun received the BEng and MEng degrees from Tsinghua University in 1986 and 1988, respectively, and the PhD degree from City University of Hong Kong, China in 2004. He is currently a professor at the Department of Computer Science and Technology, Tsinghua University. His research interests include natural language processing, social computing, Web intelligence, and computational social sciences. He has published over 150 papers in academic journals and international conferences in the above fields. He serves as a vice president of the Chinese Information Processing Society, a council member of China Computer Federation, the director of Massive Online Education Research Center of Tsinghua University, and the editor-in-chief of the *Journal of Chinese Information Processing*.

Huimin Chen received the PhD degree from Tsinghua University in 2020. She is currently a postdoctoral researcher at the School of Journalism and Communication, Tsinghua University. Her research interests include social computing and natural language processing. She has published several papers in international journals and conferences, including TBD, TKDE, ACL, EMNLP, and IJCAI.

Xuanming Zhang received the BEng degree in computer science with artificial intelligence from University of Nottingham, UK in 2020. He is currently working as a research assistant at the Department of Computer Science and Technology, Tsinghua University, China. His research interests include natural language processing and social computing.

Jianbin Jin received the BEng, BA, MEng degrees from Tsinghua University in 1991, 1992, and 1997, respectively, and the PhD degree from Hong Kong Baptist University, China in 2002. He is currently a professor at the School of Journalism and Communication, Tsinghua University. His research interests lie in the empirical studies of media adoption, uses and effects, as well as science and risk communication. He has published more than 100 papers and book chapters in academic journals and books. Currently he is a vice president of the Chinese Association of Science and Technology Communication, a council member of China Communication Association, a member of academic committee of Tsinghua University, and the executive editor-in-chief of the *Global Media Journal* published by Tsinghua University Press.