Emotional tendencies in online social networking: a statistical analysis

Xianhan Zhang\textsuperscript{a}, Nan Zhang\textsuperscript{a}, Letong Zhao\textsuperscript{a}, Ruihan Zhang\textsuperscript{a}, Jinde Cao\textsuperscript{a,\textb}, Jianquan Lu\textsuperscript{a}, Jürgen Kurths\textsuperscript{c,d,e} and Cheng Qian\textsuperscript{f}

\textsuperscript{a}Research Center for Complex Systems and Network Sciences, Department of Mathematics, Southeast University, Nanjing, People’s Republic of China; \textsuperscript{b}Department of Mathematics, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia; \textsuperscript{c}Potsdam Institute for Climate Impact Research, Potsdam, Germany; \textsuperscript{d}Institute of Physics, Humboldt University of Berlin, Berlin, Germany; \textsuperscript{e}Department of Control Theory, Nizhny Novgorod State University, Nizhny Novgorod, Russia; \textsuperscript{f}Department of Statistics, University of Michigan, Ann Arbor, MI, USA

ABSTRACT
Numerous previous studies suggested that people’s emotional tendency (ET) towards an issue can often be affected by others. But in some cases, people are unwilling to believe opposite points. This paper aims to study whether people’s emotional tendencies (ET) are susceptible with exposures to others’ ET concerning a special topic. ET contained in 798,057 pieces of private-information-deleted Chinese Weibo posts are carefully investigated via a revised genetic algorithm, a nonlinear method. Note that nearly all of the posts are closely related to a special topic, the terrible earthquake happened in Japan, 11 March 2011. By conducting statistical analysis including coefficient calculations and hypothesis testing, this study shows that concerning this particular topic, Chinese citizens’ first impressions about Japan are solid enough to form their ET and would not be easily altered. Moreover, according to analysis and discussion, we discover that node-to-node impact is exaggerated in some theoretical information diffusion models. Instead it is actually the interaction between nodes’ properties and the spread information that matters in the process of information diffusions.

ARTICLE HISTORY
Received 14 September 2015
Accepted 8 November 2015

KEYWORDS
Weibo post; genetic algorithm; emotional tendency; statistical analysis; social networking

Introduction

Online social networking fills up people’s everyday life in the twenty-first century, affecting everyone’s life through information exchanges. The fact that people continuously receives and posts in there gives rise to information (Gruhl, Guha, Liben-Nowell, & Tomkins, 2004) even rumour spread (Doerr, Fouz, & Friedrich, 2012). Famous social networking in the world provide perfect conditions for information diffusion. Various studies (Bakshy, Rosenn, Marlow, & Adamic, 2012; Cha, Mislove, & Gummadi, 2009; Lerman & Ghosh, 2010) concerning information spread in complex networks have begun since last century and they are of vital importance because diffused information in networks such as a rumour can cause impacts in people’s everyday life and meanwhile triggers possibilities for effective emotion spread. Incidents such as the salt-buying panic happened in 2011 in China illustrate the lurid power of emotion spread. In social networking, chances are that people are easily to be influenced by other people’s attitude. While some online social networking users are independent and steadfast enough to remain mentally stable when exposed to emotive feeds by celebrities, their friends and families, some evidences (Fowler et al., 2008; Kramer, Guillory, & Hancock, 2014) show that other people may experience mental changes under this circumstance.

In term of theoretical models, Li, Cong, Wu, and Wang (2014) once studied the effect of bluffing on over-confidence in social networks. They carefully investigated the role of social network in the bluffing spreading process which could also be regarded as an emotion spreading. Kramer et al. (2014) mostly focused regular posts from famous online social networking (like Facebook) data over a long-lasting period of time to examine harbingers of emotion spread. They tested whether emotion spread does exist in Facebook by reducing the amounts of positive and negative feeds read by Facebook users and thus obtained a massive amount of data. Moreover, previously there were also other previous works regarding social network focusing on data. Lindamood, Heatherly, Kantarcioglu, and Thuraisingham (2009) tried to take advantage of released social networking data to infer undisclosed private information about individuals. Melo, Moreira, Batista, Makse, and Andrade (2014) presented statistical signs of social influences on the suicide occurrence in cities. They found sublinear scaling
behaviour between the number of the suicides and the city population. These works provide us with a very reliable approach of researching: data analysis. However, different from above-listed studies, the data discussed in this paper closely revolves a theme, the horrible earthquake happened in Japan, 2011.

Puzzles still remain that whether people’s emotional tendency (ET) tend to be influenced by others in all situations. To solve this, this paper concentrates on a massive amount of data ($N = 798,057$) from Weibo (the largest online social networking in China), stretching over a relatively short span of time (9 days). Data used in this paper comes from Conference of Complex Network 2014. Private information about the Weibo users is removed and all these data is tightly related to an extremely strong earthquake that struck East Japan in March, 2011. Each piece of Weibo data consists of basic information including the time posted, the post (including stickers that are frequently used) and, etc. When the breaking news blast out and was broadcast on social network media (just like Weibo), people in China read, reposted, and added their own comments on this issue which are later read by their fans (or they can be seen by ‘stranger’ via ‘search option’). It is worth noting that these comments are mostly emotional according to our data, possibly due to that Chinese people have complex affections towards Japan. We reckon that reposters of a post have read the post and receive ET contained in it. Generally, there are two kinds of attitudes towards the breaking news in Japan among Chinese Weibo users: some Chinese citizens expressed gloating remarks, others showed sympathy towards Japanese victims. Considering the deep history plot between Japan and China during the Second World War, the former kind of attitude is not surprising. As a matter of fact, we conduct a small range of survey among 200 Chinese citizens that we come across on street, a fraction of 22.5% of them ($n = 44$) still display antagonizing attitudes against Japan. Our survey shows that this result is mainly due to the crude violence Japan executed on China during the invasion in the war. (43 of 44 respondents chose ‘history plot’ as the main reason for their attitudes.) For our data, we want to find out whether there is concrete evidence showing that previous emotions contained in posts seen by people later can effectively influence their emotional tendencies (displayed in Figure 1). Possibilities include that one of the two ET was intensified and the other weakened (the rich gets richer and the poor get poorer theory), or previous ET cannot inadvertently affect people’s judgments concerning this extreme event in 2011. We assume that people’s attitudes (or ET) are consistent with those shown in their own posts as to the matter.

In this paper, a valid and accurate method of handling regular posts in social networks is firstly designed via the combination of artificial work and an nonlinear method, genetic algorithm (GA) which is later used to judge hidden ET contained in Weibo posts. This method later proves to be of high accuracy and provides conveniences for future data-processing studies. Then, statistical tools including several well-established statistics and hypothesis testing are utilized to analyse the data. Discussions and analysis about the result is conducted. In addition, based on the results, we conclude that for this extreme event, People are heavily affected by preoccupations, and unwilling to alter their points despite receptions of opposite views.

Results

The substantial Weibo posts were finally divided into hundreds of 10-minute-long segments, the proportions of two ET (positive or negative) within each one of them could be easily judged and calculated via approaches discussed later in Method. Consequently, these Weibo posts turns into time-sorted data, made up with two proportions of the two opposite ET.

Note that the hypothesis that the emotion does spread during the studied period of time holds, iff ET is correlated with the time. Apparently, the target turns into a statistical problem. If we enclose these data into a two dimensional contingency table (Andersen, 1991), we could make use of several well-established coefficients...
including Kendall (Lapata, 2006), Gamma (Davis, 1967), Somers (Newson, 2002), Spearman (Gautheir, 2001) to test statistical correlations between the time and ET. Kendall, Gamma, and Somers are used to determine how correlated are the row and column properties in a two dimensional contingency table. Spearman’s rank correlation coefficient quantifies the statistical dependence between two variables as well, but is free from the two-dimensional contingency table. General results are listed below in Table 1.

The statistical results are displayed in Tables 1 and 2 and Figure 2–4. Firstly, nearly all of these statistics are

**Table 1.** displays analytic results (3.11–3.19) of the four statistics, Somers $d_{A,B}$, Kendall $\tau$, Gamma $\gamma$ and Spearman $\rho$.

| Time            | Somers $d_{A,B}$ | Somers $d_{B,A}$ | Kendall $\tau$ | Gamma $\gamma$ | Spearman $\rho$ |
|-----------------|------------------|------------------|----------------|----------------|----------------|
| 3.11 (15:00–17:00) | -0.006467283     | -0.023820894     | -0.012411948   | -0.025142117   | 0.027          |
| 3.12 (09:00–19:00) | -0.003278107     | -0.011875897     | -0.006239428   | -0.012077455   | 0.0543         |
| 3.13 (10:00–20:00) | 0.001478124      | 0.006111469      | 0.00300555     | 0.006214536    | -0.3213        |
| 3.14 (09:00–20:00) | -0.004698957     | -0.019111377     | -0.009476473   | -0.019520173   | 0.1163         |
| 3.15 (09:00–17:00) | -0.001148821     | -0.004624754     | -0.002304997   | -0.00472329    | 0.0967         |
| 3.16 (09:00–17:00) | 0.00147163       | 0.006269312      | 0.00303745     | 0.006372525    | -0.4547        |
| 3.17 (10:00–20:00) | 0.00147163       | 0.006269312      | 0.00303745     | 0.006372525    | -0.4547        |
| 3.18 (09:00–21:00) | 0.00147163       | 0.006269312      | 0.00303745     | 0.006372525    | -0.4547        |
| 3.19 (09:00–17:00) | 0.00147163       | 0.006269312      | 0.00303745     | 0.006372525    | -0.4547        |

Notes: As is shown in Table 1, though low correlations in three days Spearman $\rho$ displays $(0.3 < \rho < 0.5)$, other coefficients are extremely close to zero, indicating nearly no correlation between the time and ET.

**Table 2.** Additionally, in order to determine whether coefficients in Table 1 are little enough to conclude the uncorrelation between the time and emotional efficiency, hypothesis testing is conducted and chi-square $p$-values are examined.

| Time            | $\chi^2$ | $p$-Value | Time            | $\chi^2$ | $p$-Value |
|-----------------|----------|-----------|-----------------|----------|-----------|
| 3.11 (15:00–17:00) | 0.0119   | 0.9006    | 3.12 (09:00–19:00) | 0.0011   | 0.9735    |
| 3.13 (09:00–20:00) | 0.0182   | 0.8928    | 3.14 (10:00–20:00) | 0.0003   | 0.9871    |
| 3.15 (09:00–17:00) | 0.0023   | 0.9618    | 3.16 (09:00–17:00) | 0.0001   | 0.9909    |
| 3.17 (10:00–20:00) | 0.0033   | 0.9693    | 3.18 (09:00–21:00) | 0.0001   | 0.9986    |
| 3.19 (09:00–17:00) | 0.0039   | 0.9504    |

Notes: And consequently these coefficients are significantly large ($p > \alpha$, $\alpha$ is the significance, usually set to be 0.05 or 0.1). Thus the null hypothesis that no correlation exists between the time and ET is undoubtedly accepted. This result obviously strengthens the conclusion yielded from Table 1.

**Figure 2.** Nine subplots respectively plot the percentage both ET take up within each day. Blue triangle line represents the ratio of the negative ET, red snow line the positive one. Each marker in the subplot represent the result in a ten-minute long data. First subplot standing for the analytical result of the first day is incomplete since the earthquake happened in the afternoon. We can clearly tell from the plot that the proportions of the two ET stay almost constant from 3.11 to 3.19, 2011. In other words, Weibo users’ perspectives seem not susceptible to others’ ET.
Figure 3. This figure plots half of the difference between proportions of the two ET within the data time range. As is shown in each subplot, despite the fact that these lines are not perfectly horizontal, variations among days are negligibly little and this also leads us to conclude the independence between the time and the ET.

Figure 4. This figure is the candle plot displaying the evolving trend of Weibo users’ ET. The whole-time-range data is divided into forty 2-hour-unit one. Each ‘candle’ (rectangle) represents variations of ET within each time period and maximum and minimum are denoted separately by the top and bottom of the line that crosses the ‘candle’. Two fitting curves are also given on the figure and they are nearly parallel to X axis. The conclusion that Weibo users’ ET stay constant can be drawn from the relative short lengths of red, green ‘candles’ and the slight fluctuations of the two fitting curves.
extremely close to zero. Since usually we tend to believe that no correlation exists between two variables when Somers d, Kendall τ, Gamma γ and Spearman ρ are less than 0.3, the values of the four statistics are solid enough to certify the independence between the time and ET. Second, the p-values calculated from the data shown in Table 2 are much larger than the significance α (α is usually set to be 0.05 or 0.1, the calculated p-values nevertheless reach almost 1), which means the null hypothesis that the time and ET are independent cannot be rejected. Differences of the two ET in the 9 days (Figure 3) remain almost constant. What’s more, the generally steady evolving trends of the two ET and time (Figures 2 and 4) also imply the same result, because obvious rises or declines will appear in the figure if the ET of the people later see the news are evidently influenced by previous opinions. Again, analytical result indicates that ET and time are independent, neither positively nor negatively correlated. Or that is, Weibo users’ ET concerning the violent earthquake happened in Japan are not evidently susceptible to others’.

Analysis and discussion

Every quick diffusion of rumour in social network seems to prove people’s low immunity to others’ points. The reason why incidents such as ‘salt panic’ happened in China could spread so fast and caused a great amount of people to believe is that these people knew nothing about the subject. On the one hand, rumour spread can also be regarded as an emotion diffusion; people’s fears, worries about unknown staffs could effectively move from one to another. On the other hand, when the subject of the spread information is so familiar to people that they even bear preoccupations about it, chances are little that they would get strongly affected by the opposite ET. There is no denial that most Chinese citizens have previously formed their ET about Japan, the main subject of the topic, and this in a large sense lead to their unwillingness to change their stand. Then it is not difficult to understand why the evolving trend of both positive and negative ET stay stable from the very beginning to the end.

For special extreme event like this one, people who read the broadcast news are clearly influenced by preoccupations instead of ET displayed by users who previous read it. Unlike Adam et al.’s work which focused on regular everyday life Facebook post, the data discussed in this paper has a theme, a horrible earthquake happened in Japan. On the one hand, the majority of Chinese citizens were not holding an antagonizing attitudes towards Japan according to statistical result. Given the bright sides of Japan such as their positive national spirits, and technological achievements in industrial fields. On the other hand, unforgettable historical experience in the Second World War would be very likely to plant negative stereotype impression about Japan in part of Chinese citizens. Due to the old saying that ‘first impressions are strongest’, both kinds of emotions are stable or unchangeable: for people who do not possess a negative perspective about Japan, upon reception of previous unfriendly Weibo posts written by others, they are inclined to reckon that those negative comments are biased, and insist on their own points. The same goes to people who antagonize Japan, without any knowledge of Japanese virtues, they may even turn angrier when receiving previously positive ET, since they tend to hold that all Chinese citizens ought to depreciate the crime Japan conducted and should not show sympathies towards Japan. In this event, a greater fraction of the positive ET and a relatively less fraction of the negative does not result in the so-called ‘the rich gets richer and the poor gets poorer’ phenomenon. Consequently, in spite of some of these people’s exposures to opposite past ET, their preoccupations are very likely to end up even stronger and prevent them from altering to the opposite direction. Namely, the interaction between the spread information and people’s past relevant experience is the determinants factor for people’s attitude towards the spread matter.

The low effects neighbours have on an innocent node is illustrated by the statistical independence between ET and time. In fact proportions of people’s ET stay almost the same whether or not they receive prior ET. And this result to certain extent indicates possible weaknesses of some previous information spread models such as Gruhl et al. (2004), which concentrates on SIR virus diffusion model when studying information propagation through online blogs. SIR and SIS models are constructed based on the hypothesis that the difference between nodes are neglectable. Nevertheless in real social networks it is more or less irrational, now that we have clarified the interaction between the spread information and people’s personal experience. Just like this topic, young Chinese citizens must have different mental experience compared those born in 1960s since their growing environments obviously differs. Well-educated citizens may bear disparate thoughts compared with those with little educational backgrounds apparently due to their gaps of horizons or eyesights. Considering that individuals do not necessarily react the same to the spread information, data analysis result implies that a simultaneous consideration of both the properties of diffused information and nodes in the network is needed for future constructing mathematical models.
Method

Ethics statement: Data used in this research study was obtained by Complex Network Conference 2014 and complies with Weibo terms of use. All these data is fully available online by ‘search option’ from weibo.com. Additionally, this study was approved by the Ethics Committee of Knowlesys International Ltd.

Overall data processing procedures: Data mining techniques (Kirkos, Spathis, & Manolopoulos, 2007; Ngai, Xiu, & Chau, 2009; Witten & Frank, 2005) nowadays furnishes us with a variety of powerful tools to deal with potentially usable data. General data process in this paper consists of three main steps as follows (displayed in Figure 5).

(i) Although all of these Weibo posts revolves around the earthquake, we find that not all of them can be directly used in the research. Therefore we need to delete those null and incomplete Weibo posts, to ensure that all the utilized data is valid and closely related to the special issue.

(ii) We sequence the data in accordance with time for the convenience of our work. Actually the earthquake began at 11 March 2011, and instantaneously Chinese citizens started a heated discussions on Weibo about this calamity and the data started from that moment and ended on 19 March 2011.

(iii) The major part of the work is to deal with Weibo posts. We expect to take advantage of unsupervised

---

Figure 5. On the left is the general data processing procedures involved. On the right describes how GA works in this paper. Details of this proposed GA scheme can be found in the above-discussed Steps.
learning to judge what kind of emotion is contained in a Weibo post, and let the computer do the job. In this paper, we design an emotional judging approach based on GA to compensate for some possible weaknesses of previous works. The approach is designed for obtaining the optimal group of weights in order to determine ET in these Weibo data. A corresponding weight is added to a Weibo post when it contains a key clue (discussed later). Consequently, we attain a judgmental result (sum of added weights) for each piece of Weibo. It is considered negative when its judgmental result is less than zero, positive when larger than zero, neutral (abandoned) when equals to zero. In this way, we could successfully get the corresponding ET for every piece of Weibo post.

**Remark 1:** Beyond doubt, the result of this work largely depends on how precise we could determine emotions contained in Weibo posts. Obviously we have to design a scheme to do this. In prior works (Golder, & Macy, 2011; Guillory et al., 2011; Kramer et al., 2014), a text analysis program named Linguistic Inquiry and Word Count software (LIWC2007) was used. Basically, it depends on keywords searching and counting process in English. However, for Chinese, regardless of special usages of ironies, metaphors and other particular ways of expressions that frequently appear in Chinese expression, we have to emphasize that simple keywords searching and counting process is not suitable for this work to judge emotion tendencies. Apparently a more careful consideration of complex usages in online social networking expressions is needed. While the computer’s ability to judge hidden emotions contained in sentences is limited, human-beings could easily figure out what the author wants to convey. Therefore, along with artificial judgments, we make use of a revised GA to solve the problem.

**Emotion-judging method via GA:** We introduce the following definitions in the modified GA:

(a) Genes, individuals and population: We define a gene to be a weight of a clue from which we can speculate the ET of an author. The clue could be a keyword, a phrase or a special idiomatic usage. An individual possesses many genes. The combination of weights is all the genes of an individual. The population \( P(t) \) is constituted with all the individuals in the \( t \)th generation.

(b) Fitness function: We have a corresponding judgmental result of the data for every set of weights (genes). The accuracy for each set of genes towards the result of an artificial judgment is defined as the fitness function, \( f_i \) denotes the fitness value of the \( i \)th individual, and it can be obtained by \( f_i = R_i / T \), where \( R_i \) denotes the number of Weibo posts judged right (judged the same with artificial results) when the \( i \)th individual is used, \( T \) for the total number of Weibo posts. \( F_i \) is the optimal fitness value in the \( i \)th generation. The higher the accuracy, the higher the fitness value, and a greater chance for a generation to enter crossover and mutation.

(c) Crossover: We optimize this process by exchanging some genes in the parent population, which leads to a larger probability to bring about changes in the next generation towards the ideal solution (A generation is considered better when its optimal fitness value gets higher). This operation will largely enhance the efficiency of the analysis.

(d) Mutation: Apply the mutation operator to groups, that is: we randomly reassign a weight to a clue such that the probability of local convergence will decrease.

(e) Growth: We note that in real group revolution, the immigration of external species has an impact on the diversity of the local species. New individuals are constantly generated to mix into the population during the process, and thus this will lead to a greater chance of global convergence.

These definitions lead to the following modified GA consisting of 6 steps, the general flow chart of which is given in the right part of Figure 5.

**Step 1 (Artificial judgment):** First, two researchers in our team artificially and separately determine the emotions contained in 3000 pieces of Weibo posts. Only those Weibo posts judged the same by both researchers will be used as the initialization. In the whole process, new genes with weights are continuously added to the gene pool (the collection of genes) by us.

**Step 2 (Initialization):** On the basis of the artificially selected weights, we start by generating a population \( G \) with \( p \) individuals, each of which is randomly assigned a weight (uniform-distribution based, range: –20 to 20). The population \( G \) is used as initialization.

**Step 3 (Evaluation):** Individuals of the current population (\( i \)th) are evaluated by calculating the fitness value (defined above in Fitness function) of every individual and the optimal fitness \( F_i \) (the largest one) is gained.

**Step 4 (Crossover):** For the crossover operation, \( (1 – r)p \) (randomly selected from \( G \)) individuals and \( r \cdot p / 2 \) pairs of individuals selected with the probability \( \Pr(f_i) = f_i / \sum_{j=1}^{p} f_j \) (also from \( G \)) are added to a new population \( P(i + 1) \) (next generation).

**Step 5 (Mutation):** After that, as for the mutation operation, \( m \cdot p \) randomly selected individuals in \( P(i + 1) \),
are reassigned a new weight (also uniform-distribution based, range: $-20$ to $20$).

Step 6 (Termination): Lastly, the termination: the optimal weights for keywords (phrases and grammars) are reached iff $F_{i+1}$ (the optimal fitness of $P(i+1)$), satisfies

$$F_{i+1} - F_i < \varepsilon,$$

where $\varepsilon$ is a threshold set prior, otherwise steps 3, 4, 5, 6 will be repeated until termination condition is met.

Remark 2: As we are working on the 3000 Weibo posts, we find out that only a very few weights are added to the gene pool after the two of us finishing 1000 posts. After all, only a limited amount of expressions are frequently used by Weibo users. The artificial judgment is an indispensable part of our work, not only because it provides significant clue samples for the computer, but also due to the fact that it serves as a paradigm or standard for calculating the fitness value.

Details in the revised GA:

(a) Without referring to existing keyword data bases, the two researchers actually extract all the key clues. It is worth noting that not only positive or negative keywords are added to the gene pool, but also a lot of common fixed expressions including ironies, metaphors or any other that clearly display ET. For example in these Weibo data, there are many old sayings quoted, implying certain attitudes of the authors.

(b) A Weibo post is directly determined to be positive or negative when we come across stickers (such as a smile look or unhappy look) that could obviously reflect the authors’ ET. However, smile looks here actually tend to be used to express gloating attitudes, while unhappy looks are inclined to show sympathies towards the earthquake.

(c) When dealing with a negative sentence such as ‘I am not happy’ or ‘It is not a disaster’, we assign a heavier negative weight to privative than $adj$ or noun in the verbal cue. Take ‘I am not happy’ for instance, we assign a weight of $-20$ to ‘not’, $10$ to ‘happy’, resulting in an overall negative judgmental result $(-10)$ for this sentence.

(d) Eventually, we set $p = 100$, $r = m = 50\%$. And the termination threshold $\varepsilon$ is set to be less than $\frac{1}{3000}$. This bears an equal effect as to require $F_i$ and $F_{i+1}$ to be the same.

Remark 3: The combination of artificial judgments and utilizing GA in unsupervised learning enables us to ‘teach’ the computer to learn, to evolve towards the right direction, and the right direction is definitely correct since it is established by human-beings. Hence, the computer could make relatively accurate judgment about Weibo posts by comparing them to artificially judgmental result, in spite existences of complex language usages in Chinese Weibo users. Also, 1000 posts are randomly selected to test the accuracy of this proposed method. Eventually, taking artificial judgment as a standard, we have a correct rate larger than $90\% (\frac{902}{1000}, \ p > 0.99)$ which lays a solid foundation for later statistical analysis. What is more, this nonlinear approach could be further to dig more statistical data from online social network posts for other academic purposes or supervision of Chinese government.

Remark 4: When applying GA, we extract a number of keywords which appear frequently and assign each a random number. We deem a combination of all keywords as a individual and the numbers of keywords as the genes of the individual. The difference between of our algorithm and others is that we directly apply crossover operation and mutation operation on the numbers of keywords without transcoding them into binary system or some other code. This kind of application brings about the advantage that we have a short timer to computation.

Statistical method: After dividing the substantial amount of posts into 10-minute-unit segments, we utilize statistical tools as follows to analyse the correlation between ET and the time.

Calculating statistics and p-values: Before calculating the four coefficients, we need to introduce several frequently used auxiliaries, $G, H, T_A, T_B, T_{AB}$, which are defined as follows in (1)–(5)

\[
G = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{ij} \left( \sum_{k=i+1}^{c} \sum_{t=j+1}^{c} n_{kt} \right) = \sum_{i} \sum_{j} n_{ij} n_{kt}, \tag{1}
\]

\[
H = \sum_{i=1}^{r-1} \sum_{j=2}^{c} n_{ij} \left( \sum_{k=i+1}^{c} \sum_{t=1}^{t} n_{kt} \right) = \sum_{i} \sum_{j} n_{ij} n_{kt}, \tag{2}
\]

\[
T_A = \sum_{i=1}^{r} \left( \frac{n_{i+}}{2} \right) = \sum_{i=1}^{r} \frac{n_{i+}(n_{i+} - 1)}{2}, \tag{3}
\]

\[
T_B = \sum_{j=1}^{c} \left( \frac{n_{+j}}{2} \right) = \sum_{j=1}^{c} \frac{n_{+j}(n_{+j} - 1)}{2}, \tag{4}
\]

\[
T_{AB} = \sum_{i=1}^{r} \sum_{j=1}^{c} \left( \frac{n_{ij}}{2} \right) = \sum_{i} \sum_{j} n_{ij} (n_{ij} - 1) \frac{2}{2}, \tag{5}
\]

where $n_{ij}$ is the value of the one in row $i$ and column $j$, $n_{i+}$ denotes the sum of row $i$, and $n_{+j}$ stands for the sum of column $j$. Based on these auxiliaries, we could provide computational formulas above-mentioned the four statistics.
(1) Kendall’s τ

\[ \tau = \frac{G - H}{\sqrt{n(n-1)/2 - T_A}} \]  

where \(n\) is the sample size, \(G\) and \(H\) are the number of concordant and discordant pairs, and \(T_A\) is a measure of the number of ties in the ranks of variable \(A\).

(2) Gamma coefficient

\[ \tau = \frac{G - H}{G + H}. \]

(3) Somers’ \(d\) coefficients

\[ d_{B|A} = \frac{G - H}{n(n-1)/2 - T_A}, \]
\[ d_{A|B} = \frac{G - H}{n(n-1)/2 - T_B}, \]

where \(d_{A|B}\) and \(d_{B|A}\) respectively measure the effect property \(A\) has on \(B\) and that \(B\) has on \(A\).

(4) Spearman’s rank correlation coefficient: Another way to see how correlated two variables are is to calculate Spearman’s rank correlation coefficient \(\rho\). It is a nonparametric measure of statistical dependence between two variables, which is gained from (10):

\[ \rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \]

where \(d_i\) is the difference between ranks of \(x_i\) and \(y_i\).

What’s more, with the null hypothesis to be the ET and the time are independent and the alternative hypothesis to be the opposite, we could compute the chi-square value for hypothesis testing:

\[ \chi^2 = U^2 = \frac{(G - H)^2}{\sigma^2(G - H)}, \]

where

\[ \sigma^2(G - H) \approx \frac{(n^3 - \Sigma n_i^3)(n^3 - \Sigma n_j^3)}{9n^3}. \]

When the null hypothesis holds, \(\chi^2\) is asymptotically chi-square distributed, and its degree freedom equals to 1.

**Concluding remarks and future work**

In order to test the susceptibility of Weibo users’ ET, a highly accurate nonlinear method via GA has been devised to judge hidden ET in a substantial amount of Weibo posts, and all of these posts revolves around the horrible earthquake happened in Japan, 2011. Statistical tools including four well-known coefficients and hypothesis testing have been utilized to test correlations between Weibo users’ ET and the time. As a result, ET and the time are actually independent from each other. Therefore, Chinese Weibo users’ first impressions about Japan, no matter positive or negative, still thrive, that is, Weibo users’ ET as to this extreme event are immune to existing emotional remarks. Our next concern will be utilizing the proposed data mining method to gain more reliable real social network data to better understand spreading patterns in social network. The interaction between nodes, and the information spread will also be very meaningful. An information diffusion model that considers both properties of nodes and the diffused information may be more practical and better fit reality.

**Acknowledgments**

The statements made herein are solely the responsibility of the author[s]. The author would also like to thank Professor Hongzhong Zhang from Columbia University for his kind advice about this paper. Author contributions: X.Z., J.C., J.L. conceived the research, X.Z., N.Z., L.Z., R.Z., J.C., J.L., J.K., C.Q. designed and performed the research. X.Z., L.Z. analysed the results and X.Z. J.C., J.L., J.K., C.Q. wrote, reviewed and edited the paper. All authors read and approved the paper.

**Disclosure statement**

No potential conflict of interest was reported by the authors.

**Funding**

The authors acknowledge the National Natural Science Foundation of China under Grant 61175119 and Grant 61272530, and IRTG 1740 (DFG and FAPESP).

**References**

Andersen, E. B. (1991). Two-way contingency tables. In *The statistical analysis of categorical data* (pp. 89–130). Berlin: Springer.

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). *The role of social networks in information diffusion*. Proceedings of the 21st International Conference on World Wide Web, ACM, pp. 519–528.

Cha, M., Mislove, A., & Gummadi, K. P. (2009). *A measurement-driven analysis of information propagation in the flickr social network*. Proceedings of the 18th International Conference on World Wide Web, ACM, pp. 721–730.

Davis, J. A. (1967). A partial coefficient for Goodman and Kruskal’s gamma. *Journal of the American Statistical Association, 62*, 189–193.

Doerr, B., Fouz, M., & Friedrich, T. (2012). Why rumors spread so quickly in social networks. *Communications of the ACM, 55*, 70–75.

Fowler, J. H., Christakis, N. A. (2008). Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the framingham heart study. *BMJ, 337*, 2338.

Gauthier, T. D. (2001). Detecting trends using Spearman’s rank correlation coefficient. *Environmental Forensics, 2*, 359–362.

Golder, S. A., & Macy, M. W. (2011). Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science, 333*, 1878–1881.

Gruhl, D., Guha, R., Liben-Nowell, D., & Tomkins, A. (2004). *Information diffusion through blogspace*. Proceedings of the
13th International Conference on World Wide Web, ACM, pp. 491–501.
Guillory, J., Spiegel, J., Drislane, M., Weiss, B., Donner, W., & Hancock, J. (2011). *Upset now? emotion contagion in distributed groups*. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, pp. 745–748.
Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111, 8788–8790.
Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications*, 32, 995–1003.
Lapata, M. (2006). Automatic evaluation of information ordering: Kendall’s tau. *Computational Linguistics*, 32, 471–484.
Lerman, K., & Ghosh, R. (2010). Information contagion: An empirical study of the spread of news on digg and twitter social networks. *ICWSM*, 10, 90–97.
Li, K., Cong, R., Wu, T., & Wang, L. (2014). Bluffing promotes overconfidence on social networks. *Scientific Reports*, 4.
Lindamood, J., Heatherly, R., Kantarcıoglu, M., & Thuraisingham, B. (2009). *Inferring private information using social network data*. Proceedings of the 18th International Conference on World Wide Web, ACM, pp. 1145–1146.
Melo, H. P. M., Moreira, A. A., Batista, É., Makse, H. A., & Andrade, J. S. (2014). Statistical signs of social influence on suicides. *Scientific Reports*, 4.
Newson, R. (2002). Parameters behind ‘nonparametric’ statistics: Kendall's tau, somers’ d and median differences. *Stata Journal*, 2, 45–64.
Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36, 2592–2602.
Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. Burlington: Morgan Kaufmann.