Distance Decay Effect and Spatial Interaction during the COVID-19 pandemic

Nicolas Wolz [1], Manning Xu [2], Tiantian Wang [3]

Abstract:

In computational communication science, social network data can be used to analyze trends in the communication behavior of people. For this purpose, a data set containing English Tweets was provided by the University of Technology Ilmenau, which was collected during the beginning of the COVID-19 pandemic by the Database Systems Research Group at University Heidelberg. The goal was to find hidden patterns within the data to show if and how the pandemic influenced our communication. This paper looks at the Distance Decay Effect, which says that near things are more related to each other than distant things, and therefore communication should get more sparse the greater the distance between users. Modeling the data with a Gravity Model shows that this relationship is true for the data provided, therefore reproducing earlier research on this topic. We were not successful in finding any clear trend showing that the strength of the Distance Decay Effect changed over the course of the first weeks of the pandemic.

Keywords: Distance Decay Effect; Gravity Model; COVID-19; Twitter

1 Introduction

An increasing demand for understanding the spatial connection characteristics including spatial and temporal perspectives of information diffusion on social media is observed in various practical scenarios [Su17]. For instance, the government needs to contain the diffusion of rumors on social media, to identify critical geographic areas and key time windows where rumors originate from, become viral and fade out, which can further help them disseminate the truth to users in those critical areas and at optimal time spots [Ca12] [SOM10] [Su17]. Some research shows that investigating the relationship between cyberspace and real space, using big data and social media data, can help better understand human activities [HTC18] [Ju15].

Twitter has become a feasible social media platform to explore global human communication patterns as well as city-scale human communication. It has been widely used as a tool

[1] TU Ilmenau, Student / Databases and Information Systems Group, Ehrenbergstraße 29, 98693 Ilmenau, Germany nicolas.wolz@tu-ilmenau.de
[2] TU Ilmenau, Student / Department of Economic Sciences and Media, Computational Communication Research Group, Ehrenbergstraße 29, 98693 Ilmenau, Germany manning.xu@tu-ilmenau.de
[3] TU Ilmenau, Student / Department of Economic Sciences and Media, Computational Communication Research Group, Ehrenbergstraße 29, 98693 Ilmenau, Germany tiantian.wang@tu-ilmenau.de

to understand group dynamics from information dissemination on online social networks [LG10]. During the situation of the COVID-19 pandemic, more importantly, Twitter provides the opportunity for researchers to explore the role social media plays, especially in a global health crisis [CE10]. Thus, Twitter has become one of the centers of the social infrastructure and is a technology that allows us to stay connected, even during the crisis [CLF20].

COVID-19 has been characterized as a pandemic on March 11. At that point, the virus has affected 114 countries, which led to unnecessary suffering and death [WH]. Some measures like social distancing, quarantines, travel bans, and business closures are changing the structure of societies worldwide. Due to being forced out of public spaces, a lot of people communicate about the pandemic in social media, like Twitter [CLF20]. The numbers of the conversation around COVID-19 have continued to expand [Ab20]. However, the influence of physical distance in digital interaction is still a topic of research [HTC18]. Some researchers like Han, Tsou, and Clarke [HTC15] find that the individuals who live in nearby places interact more with each other than the people living more distant from each other. This phenomenon is called Distance Decay Effect. To describe this effect, researchers use the gravity model [Yu17], which describes the expected rate of interaction given a distance between places, based on the data the model was trained on.

This paper aims to find the Distance Decay Effect during the breakout time of COVID-19 pandemic, using the Gravity Model and Twitter data from March 2020, extracted by the Database Systems Research Group at University Heidelberg [4]. Moreover, we want to investigate whether the COVID-19 pandemic changed this communication pattern.

2 Theoretical Considerations & Research Questions

2.1 Theoretical background

Spatial connection was implied in Tobler's first law of geography (Tobler, 1970), the concept is "near things are more related than distant things", and covers a broader range of connection than "interaction" [Yu17]. Extant research gave spatial interaction different definitions. MacLachlan [Qu] defined spatial interaction as a dynamic flow process that articulates one location with another. It is a general concept that may refer to the movement of human beings such as intra-urban commuters or intercontinental migrants but may also refer to traffic in goods such as raw materials or to flows of intangibles such as information. Other researchers define spatial interaction in a broad context as actual or potential flow among places, with any type of connection among places [HTC18].

During the COVID-19 pandemic, people produced social media content on Twitter. The place of origin and the coordinates of users' locations while posting are recorded and

[4] https://dbs.ifi.uni-heidelberg.de

are partly available for research. It is also recorded if a tweet is a reply to another tweet. Therefore, in the context of this paper, a spatial interaction is defined as a tweet replying to another tweet. The distance of this interaction is calculated using the coordinates (latitude, longitude) associated with the respective tweets. The whole study revolves around the question how many people from inside a so-called central entity interact with the outside world and vice versa. This central entity is defined as the set of tweets posted within a given radius of a central coordinate. We chose the city centers of several major US cities (New York, Los Angeles, Seattle, Chicago, Houston) as central coordinates and 100 km radius to include the whole metropolitan area of those cities. Tweets from outside the central entity are grouped in "distance from this central city"categories. By looking inside the US on a city scale, we have a large language-homogenous area for the distance decay effect to occur with less language bias than observed on a global scale.

In order to analyze the data, the commonly used Gravity Model is leveraged. It is used in several research with regard to spatial interaction due to its effectiveness in predicting the degree of interaction, the simplicity of its equation, and its ability to deal with flows in both directions [HFG12]. Distance decay is inherent in spatial gravity models, but the slope and range of the distance decay vary depending on the type of human interaction [HTC18]. The model can be applied as a qualitative conceptual tool or it may be operationalized in different ways using quantitative data [Qu]. We shall employ a simple bivariate regression model to estimate model parameters to quantify the distance decay effect on interaction.

2.2 Study design and research questions

In general, this paper investigates how the COVID-19 Pandemic affected people's communication behavior. To answer this question, the study analyzes the provided Twitter data set in an observational approach. Two research interests were developed to structure the research.

Research interest 1: If the distance from the central entity is shorter, the amount of spatial interaction is higher.

Based on the research of distance decay and spatial interaction by Han, Tsou, and Clarke [HTC15], our research interest is phrased as a hypothesis. This part of the research is conducted confirmatory.

Research interest 2: We will continue to approach the data explorational to see if there has been a change in the spatial interaction behavior over time. Temporally, this research focuses on data collected in March. During this month, both Europe and the US have seen a strong increase in COVID-19 cases and some of the most important events of the pandemic happened in this time period in respective calendar weeks (CW):

It could be expected that the strength of the Distance Decay Effect changed during the course of the COVID-19 pandemic, especially with lockdowns enforced all over the word. This research aims to explore if any general trend can be revealed.

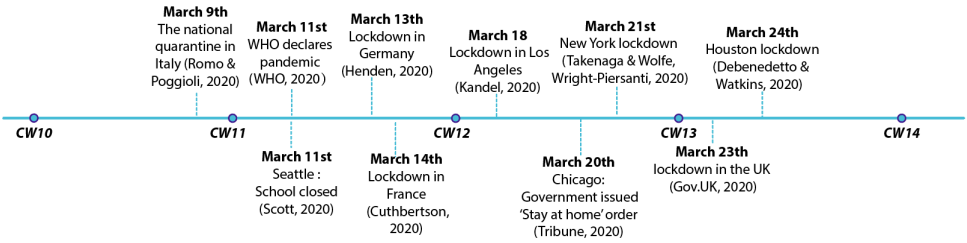

Fig. 1: Timeline of the lockdown time in European countries and five US cities.

3 Methodology

This chapter focuses on the main ideas and the model used in order to be able to understand the conclusions drawn from the results. First, the definition of terms, denotations and observed variables will be introduced. Secondly, the process of finding a Distance Decay Effect in the data is split in two distinct steps: data filtering and information extraction; building, evaluating and visualizing of Distance Decay Effect and Gravity Model.

3.1 Definition of terms, denotations and observed variables

As mentioned, a spatial interaction is defined as a tweet replying to another tweet. In order to calculate the number of those interactions, the attributes “ID” and “in_Reply_to_status_id” of the given Twitter data set are used. The number of spatial interactions is denoted as I_ij (I:big letter i).

The central entity is the middle point of our model. Like already mentioned, it represents the metropolitan area of a city. To find out if a Tweet belongs to the central entity, the coordinates found in the attribute “bounding_box” are used. Model variables belonging to the central entity are denoted with the index i. On the other hand, the outside world will be denoted with index j. Tweets get categorized based on the distance from the central entity.

The distance between the central entity i and a place in the outside world j is calculated using the coordinates found in “bounding_box”. For this, the haversine formula [Ha] is used to calculate the distance between two points on a sphere with given longitudes and latitudes. It is denoted as D_ij. The conceptual size of a group is the total number of tweets in it. This measurement is used to normalize the data, since it is obvious that the pure size of some places, like Los Angeles, will increase the number of interactions. In order to remove this effect from the results, normalization is needed. The letter P is used as a denotation for the conceptual size.

3.2 Data Filtering and Information Extraction

The first step consists of reading the big data set, which contains about 100 Million Tweets, including attributes mentioned in the previous subsection, among others. It gets condensed down to a much smaller data sample with a higher information density, containing a small number of key observations. For that, the source data is iterated two times. This results in a small .json file containing all the relevant information needed for the second step, data analysis (see next subsection).

The first iteration is used to extract tweets originating from the central entity within a given time frame. The items of columns "ID" (set A) and "in_Reply_to_status_id" (set B) are stored in sets in order to work with them during the second iteration.

During the second iteration, each tweet is categorized/grouped, based on its distance from the central entity. In order to find spatial interactions, according to the definition corresponding to replies, the items of columns "ID" (set C) and "in_Reply_to_status_id" (set D) of each category are stored in sets as well. The number of interactions is calculated by adding $|A \cap D|$ and $|B \cap C|$. Additionally, the conceptual sizes of each group gets counted as well. After some data transformation, the output of this step looks like shown in figure 2 (note that just the head of the data frame is shown here, the complete data frame contains more entries).

	conceptualSize	spatialInteractions	distance
100	757716	91	100
200	826967	64	200
300	1354098	131	300
400	503594	30	400
500	388781	13	500

Fig. 2: Structure of the intermediate results

3.3 Data Analysis

3.3.1 Distance Decay Effect and Gravity Model

The second step consists of building, evaluating and visualizing the gravity model. As described in Yuan, Liu and Wei's [Yu17] research, this is a simple and effective method to model the Distance Decay Effect based on observations.

The observations, which are the output of the first step (see figure 2) and thus the input for this step, consist of the number of spatial interactions I_ij between our central entity i and

$$I_{ij} = K\frac{P_i P_j}{D_{ij}^{\beta}} \tag{1}$$

Equation 1: Gravity model formula

the outside world places j, the respective conceptual sizes P_i and P_j of those places, and the distances D_ij between them.

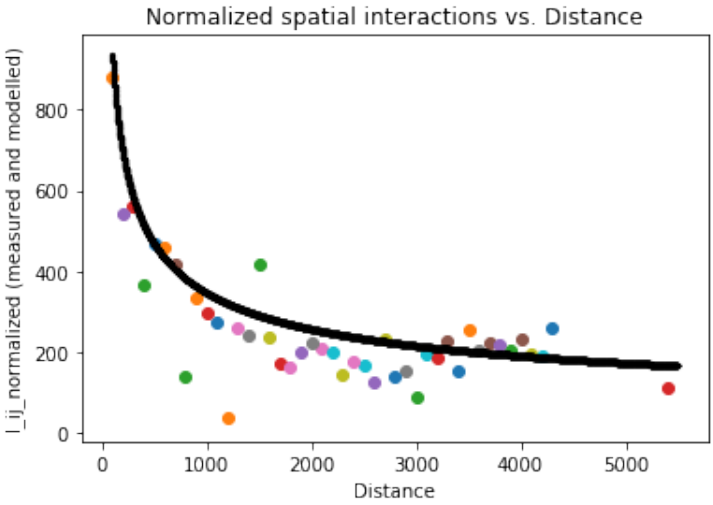

Fig. 3: Gravity model curve and observed data points

As seen in figure 3 and figure 4, we can plot the gravity model curve and a correlation graph between modelled and observed data, if we find values for the constant K, which is a scaling factor, and for β, which is the distance friction coefficient. In order to determine those two missing values, we do a so-called model fitting.

3.3.2 Model fitting and evaluation

To find the distance friction coefficient β, we iterate through a lot of potential values (e.g. 0 to 3 in very small steps) and evaluate each resulting model (see figure 5). The β with the highest evaluation score, called R_squared, will be chosen for our final model. R_squared is calculated by squaring the Pearson Correlation between the model and the observation (see figure 4). The β with the highest score is called "best matching β".

Since K is a constant scaling factor and does not influence correlation and therefore the quality of our model, a rough estimation of a realistic value is sufficient. It would be possible

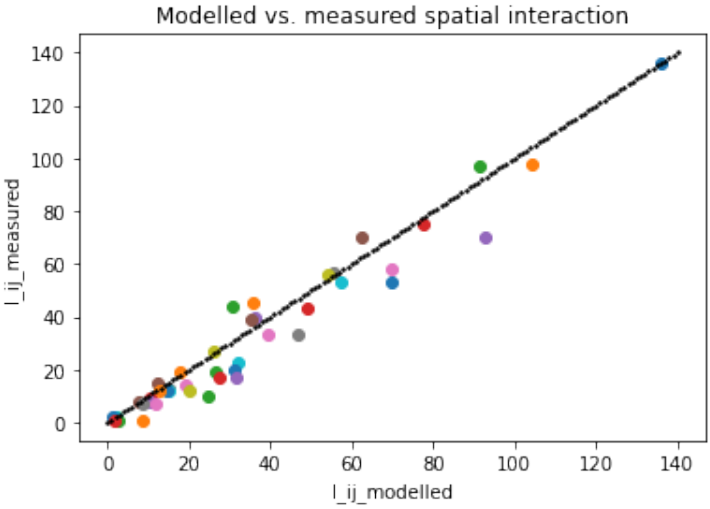

Fig. 4: Correlation between measured and modelled spatial interaction

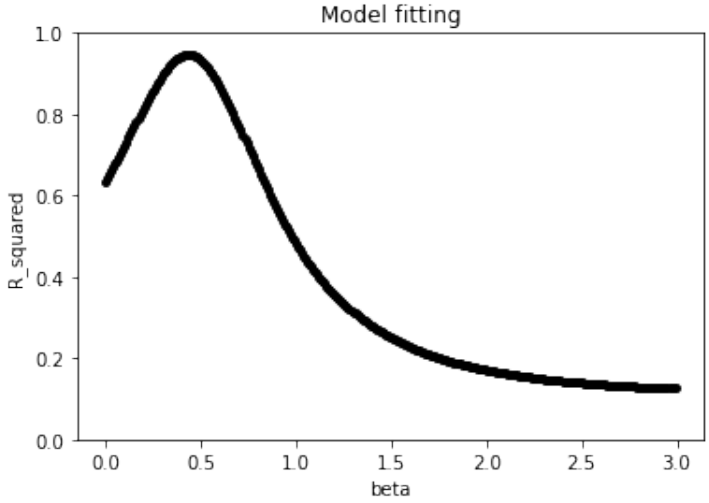

Fig. 5: Iteration over many values for β and evaluation of each resulting model

to do another round of model fitting, this time with fixed β and variable K, using e.g. least square optimization, to find the perfect value. For the purpose of this research this was not necessary, since only the best matching β and its respective R_squared value are relevant. A realistic value for K is just needed for plots like figure 3 and figure 4. For those, it is calculated using the values of just one data point in order to have a rough estimation, which is not necessarily the best fitting value.

4 Results

As mentioned before, the central entities represent the metropolitan areas around a central coordinate in the middle of the major US cities of New York, Los Angeles, Seattle, Chicago and Houston. For each city and time frame, a separate model is being built.

Regarding research interest 1, analysis was conducted using data from the whole month of March. The model fitting process produces the values seen in figure 6. The Gravity Model produces the expected curves and correlation diagrams as seen in figure 3 and figure 4.

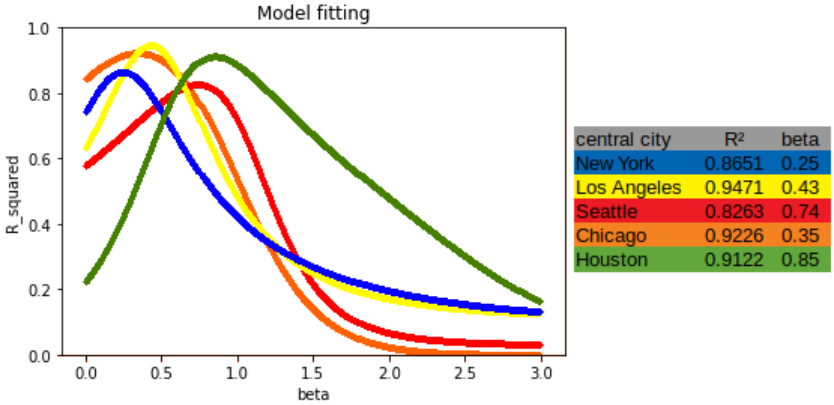

Fig. 6: Model fitting for five major cities in the US

Therefore, a distance decay effect inside the US can be clearly observed. If the hypothesis of research interest 1 was false, the results would show very low or unrealistically high values for the correlation R_squared or unrealistic values for β, e.g. 0.

In order to explore research interest 2, the data was analyzed for specific weeks (CW10 to CW13). As it can be seen in figure 7 (right), R_squared improved significantly from CW10 to CW11 for every city, which is in line with the growth of the number of tweets of each city present in the data set for those weeks. However, this does not mean that a high number of tweets automatically results in a higher R_squared score since this correlation does not continue in the following weeks CW12 and CW13.

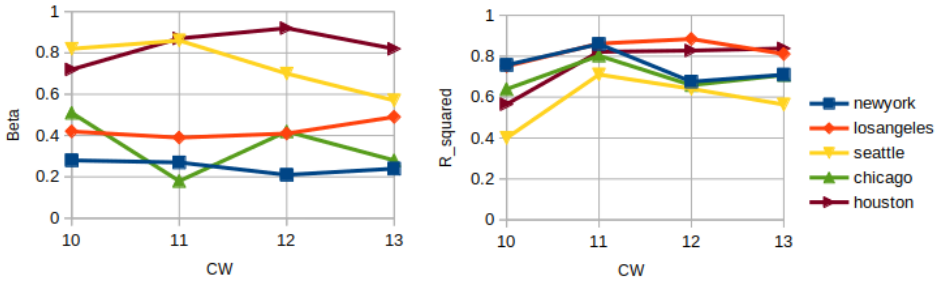

Fig. 7: Development of β and R_squared for CW10 to CW13

Figure 7 (left) shows the development of β. New York and Los Angeles only show a slight fluctuation within this time frame. Besides that, Chicago reached a low β value in CW11, shortly before the lockdown time (March 20th). However, the values for Chicago generally show a great fluctuation. Furthermore, Seattle indicates a downwards trend from CW11 to CW13, while Houston shows an upward trend from CW10 to CW12.

Overall, there is no uniform and clear trend to be seen for all cities for the value of β, given our data set and methodology. This could mean that the strength of the distance decay effect was not affected by the events of the pandemic during this time frame in the US.

5 Discussion

Initially, this study tried to analyze the data on a global scale with whole countries as entities in mind. Unfortunately, this approach was not successful because of limitations which will be discussed in the next subsection. However, with the adapted US city approach, the findings are in line with Han, Tsou, Ming-Hsiang, Clarke's (2018) [HTC18] research. As is notable in figure 8, there is a high correlation (0.98) between the results of this research and theirs. However, the definition of spatial interaction and therefore the absolute numbers differ. In Han, Tsou, Ming-Hsiang, Clarke's (2018) [HTC18] research, "following" was one way to measure spatial interaction. In this research, only "replying" was considered spatial interaction.

As for research interest 2, the results do not meet the previous expectations, as no correlation between Distance Decay Effect and the events of COVID-19 pandemic was found. As mentioned, this might indicate that there is no connection and the communication patterns were not influenced by the pandemic. But it might also be the case that the limitations, which will be discussed in the next subsection, are the reason for absence of evidence. Exploring the development of spatial interaction over a longer period of time, using a bigger and possibly better data set and defining more means of spatial interaction could help to reveal those trends and correlations, if they exist.

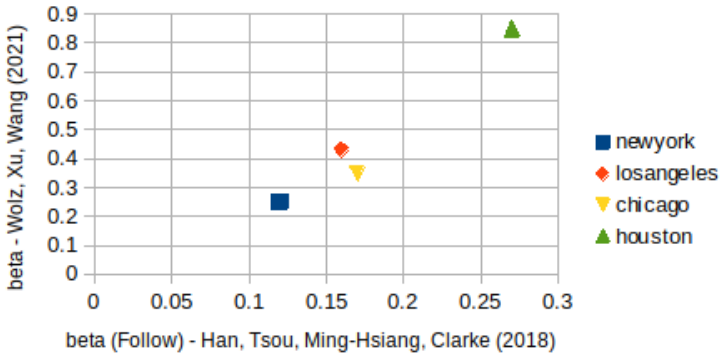

Fig. 8: Correlation (0.98) between our findings and Han, Tsou, Ming-Hsiang, Clarke (2018) [HTC18] [Figure 7, "Follow"]

5.1 Limitations

Regarding limitations, one of the main problems of this study was the strong language bias present in the data set. Since the data set contains only tweets labelled as English language, communication in and between non-English speaking countries is very sparse. The countries Great-Britain, United States, Ireland and Canada dominate the data set and therefore the model fitting process. Since model fitting is done using linear correlation between modelled and measured interactions, the data points with the biggest number of interactions strongly dictate the results, as seen in figure 9. Based on this finding, we conclude that it is not possible to research the Distance Decay effect on a global-scale, using our methodology and the given English-biased data set. In order to conduct more research on this topic on a global scale, a non-biased data set has to be collected.

Another limitation of this research is the definition of spatial interaction. Since only replies to a tweet are considered, other means of interaction are ignored. Besides the obvious liking and retweeting, also physical movement or mentioning of places could be considered spatial interaction [HTC18].

For some models, the R_square values are quite low (see figure 7, CW10). In some cases, this correlates with a low number of tweets and therefore a low number of spatial interactions, which makes the performance of the gravity model more susceptible to noise and randomness.

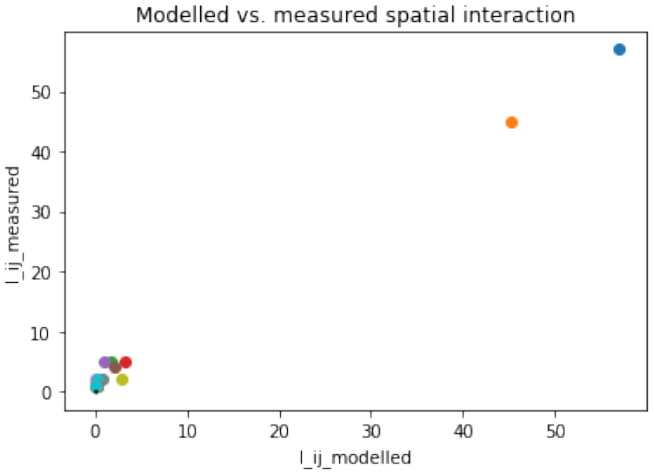

Fig. 9: Correlation graph for central entity Germany for the time from CW10 to CW13

5.2 Outlook

Further research on this topic could try to tackle mentioned problems in order to verify the findings of this research or find new trends and effects. The general methodology can be used on other data sets, possibly containing also non-English Tweets. Complementing the methodology with further definitions of spatial interaction and analyzing a longer time scope might also help to produce new insights.

6 Acknowledgements

We would like to thank both the University of Technology Ilmenau and the Heidelberg University for making this paper possible. The presented results have been achieved during the research seminar „Data Science Social Media“, supervised by Aliya Andrich, Prof. Emese Domahidi, Prof. Kai-Uwe Sattler and Dr. Nadine Steinmetz. The Twitter data set has been retrieved with the help of the Database Systems Research Group at Heidelberg University.

Bibliography

[Ab20] Abbas, Ansar; Eliyana, Anis; Ekowati, Dian; Saud, Muhammad; Raza, Ali; Wardani, Ratna: Data set on coping strategies in the digital age: The role of psychological well-being and social capital among university students in Java Timor, Surabaya, Indonesia. Data in Brief, 30:105583, 2020.

[Ca12] Cao, Nan; Lin, Yu-Ru; Sun, Xiaohua; Lazer, David; Liu, Shixia; Qu, Huamin: Whisper: Tracing the Spatiotemporal Process of Information Diffusion in Real Time. IEEE Transactions on Visualization and Computer Graphics, 18(12):2649–2658, 2012.

[CE10] Chew, Cynthia; Eysenbach, Gunther: Pandemics in the Age of Twitter: Content Analysis of Tweets During the 2009 H1N1 Outbreak. PloS one, 5:e14118, 11 2010.

[CLF20] Chen, Emily; Lerman, Kristina; Ferrara, Emilio: Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. JMIR public health and surveillance, 6(2), 2020.

[Ha] Haversine formula. https://pypi.org/project/haversine/. Accessed: 2020-07-21.

[HFG12] Hardy, Darren; Frew, James; Goodchild, Michael F.: Volunteered geographic information production as a spatial process. International Journal of Geographical Information Science, 26(7):1191–1212, 2012.

[HTC15] Han, Su Yeon; Tsou, Ming-Hsiang; Clarke, Keith C.: Do Global Cities Enable Global Views? Using Twitter to Quantify the Level of Geographical Awareness of U.S. Cities. PLOS ONE, 10(7):1–23, 07 2015.

[HTC18] Han, Su Yeon; Tsou, Ming-Hsiang; Clarke, Keith C.: Revisiting the death of geography in the era of Big Data: the friction of distance in cyberspace and real space. International Journal of Digital Earth, 11(5):451–469, 2018.

[Ju15] Jurdak, Raja; Zhao, Kun; Liu, Jiajun; Abou Jaoude, Maurice; Cameron, Mark; Newth, David: Understanding Human Mobility from Twitter. PLoS ONE, 10, 07 2015.

[LG10] Lerman, Kristina; Ghosh, Rumi: , Information Contagion: an Empirical Study of the Spread of News on Digg and Twitter Social Networks, 2010.

[Qu] Quantitative Models for Geographical Analysis. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.471.6867&rep=rep1&type=pdf. Accessed: 2020-06-11.

[SOM10] Sakaki, Takeshi; Okazaki, Makoto; Matsuo, Yutaka: Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors. In: Proceedings of the 19th International Conference on World Wide Web. WWW '10, Association for Computing Machinery, New York, NY, USA, S. 851–860, 2010.

[Su17] Sun, Guodao; Tang, Tan; Peng, Tai-Quan; Liang, Ronghua; Wu, Yingcai: SocialWave: Visual Analysis of Spatio-Temporal Diffusion of Information on Social Media. ACM Trans. Intell. Syst. Technol., 9(2), Oktober 2017.

[WH] WHO Emergencies Press Conference on coronavirus disease outbreak-11 March 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/media-resources/press-briefing. Accessed: 2020-03-11.

[Yu17] Yuan, Yihong: Exploring the Spatial Decay Effect in Mass Media and Location-Based Social Media: A Case Study of China. In (Griffith, Daniel A.; Chun, Yongwan; Dean, Denis J., Hrsg.): Advances in Geocomputation. Springer International Publishing, Cham, S. 133–142, 2017.