Data Reliability Assessment based on subjective opinions

by

Danchen Zhang

Master, Beihang University, 2011 - 2013

Bachelor, Nanchang University, 2007 - 2011

Submitted to the Graduate Faculty of

the School of Computing and Information in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

SCHOOL OF COMPUTING AND INFORMATION

This dissertation was presented

by

Danchen Zhang

It was defended on

May 18, 2021

and approved by

Vladimir I Zadorozhny, School of Computing and Information, University of Pittsburgh

Daqing He, School of Computing and Information, University of Pittsburgh

Vladimir A Oleshchuk, Department of Information and Communication Technology, University of Agder

Kostas Pelechrinis, School of Computing and Information, University of Pittsburgh

Dissertation Director: Vladimir I Zadorozhny, School of Computing and Information, University of Pittsburgh

Data Reliability Assessment based on subjective opinions

Danchen Zhang, PhD

University of Pittsburgh, 2021

In the big data era, numerous data fluctuates society and people's life. These data come from diverse sources, and various information can be inferred and extracted. However, data quality usually cannot be guaranteed, and hence decision making with such unreliable data may lead to considerable losses. Accurate data reliability assessment mechanisms can help recognize the distrustful information and then filter unreliable data out.

In this work, I consider a novel approach to assess data reliability based on subjective opinions. I structure the data propagation model in terms of data sources producing and evaluating different statements. Next, I explore data history labels, value conflicts, and uncertainty. For different combinations of those parameters, I consider common scenarios, including handling fake news, truth discovery, data cleaning, as well as discovering cancer-driving genes.

In my dissertation, I explore how to accurately assess data reliability and how to make a decision based on evaluated reliability. I propose a series of subjective opinion based models to assess each scenario's reliability and compare them with state-of-art models through experiments on real-world data.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

In this chapter, the background and motivation of this study are discussed in the first section, including different data reliability assessment scenarios and Subjective Opinion based data reliability representations. The problem statement is presented in Section 2, and then I go through four research questions in Section 3. Lastly, I outline and summarize the remaining work of this dissertation.

1.1 Background and Motivation

Data reliability has always been considered as a very important issue. Data come from diverse sources, fluctuating our daily lives. People make decision based on the information obtained from such data. However, data quality usually cannot be guaranteed, and hence decision making with such unreliable data may lead to loss beyond estimation. For example, in the 2016 US presidential election season, fake news about Hillary Clinton fluctuated social media and had an indispensable effect on the election results. Accurate data reliability assessment mechanisms can help people or machines recognize the distrustful information and then filter unreliable data out. Afterward, people can make decisions with clean, consistent, and reliable information.

In this study, I focus on the data reliability assessment and assume that data sources provide values for various statements. Data sources may be unreliable, providing wrong values, and also, the statement may be fake. This dissertation study aims to assess data reliability accurately.

A **Statement** is composed based on an object or a fact. For example, given a restaurant, a statement may be ”this restaurant is pricy”; given the fake news that Avril Lavigne died in 2003, a statement could be ”this news is real.” When we doubt the correctness of the statement, we say it has reliability issues.

Meanwhile, **data sources** provide **values** to statements, reflecting the data source’s

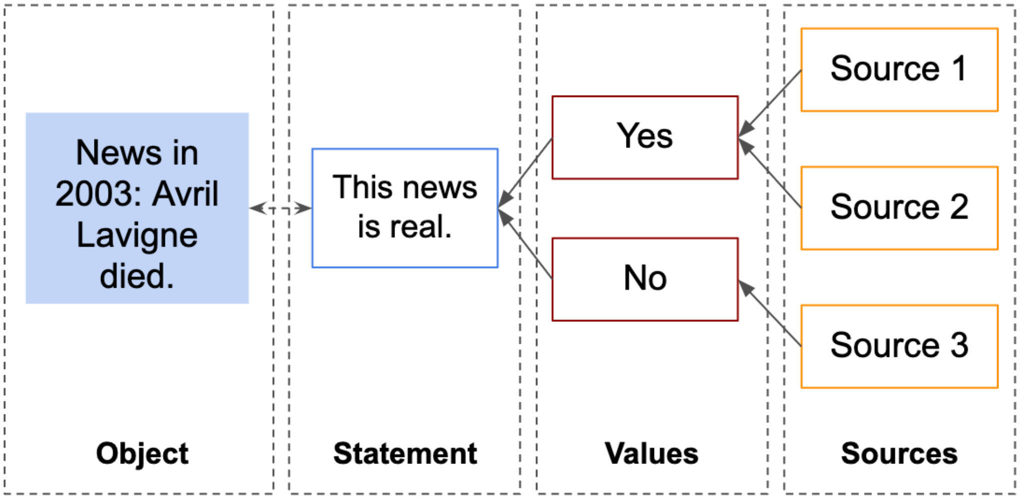


Figure 1: Three sources provides different values to a statement.

opinion or stance, as shown in Figure 1. For example, source A says ”Yes, Avril died in 2003”, source B says ”No, she died in 2011”, and source C says ”No, she is alive”. The values of a statement could be conflicting. It indicates that some data sources are providing wrong values, and we say the data sources providing wrong values are unreliable. Also, among the candidate values, we want to figure the most likely correct values.

Figure 1 illustrates the scenario with only one statement. In real-world scenarios, there are usually many statements, and many data sources simultaneously providing values to them, constructing a much more complicated network, as shown in Figure 2.

To sum up, in this study, I explore the following problems: (1) given multiple data sources, which of them are reliable; (2) given the statements, which statements are correct; (3) given conflicting values, which value is true. I address these questions to assess data reliability in different scenarios, as explained in the next section.

1.1.1 Different data reliability assessment scenarios

In this section, I identify four different scenarios based on whether historical data is provided or not, and whether values are conflicting, as shown in Table 1.

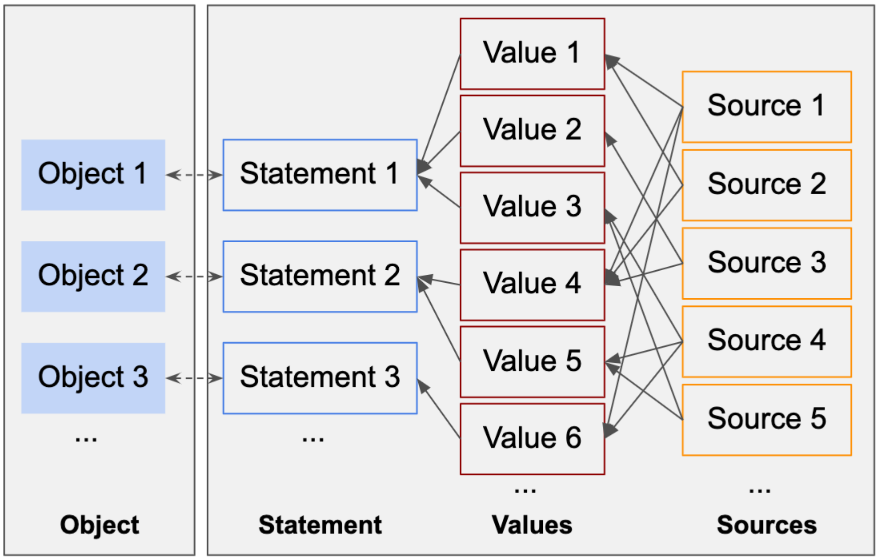


Figure 2: Different sources provide different values on multiple statements.

Table 1: Four different scenarios for data reliability assessment, and corresponding real-world problems in this study.

	No historical statement labels	**Has historical statement labels**
Single value	Scenario 1. E.g., find cancer driver genes.	Scenario 2. E.g., fake news detection
Multiple values	Scenario 3. E.g., find true book author list	Scenario 4. E.g., find true city populations

In some scenarios, historical statement labels are provided. For example, with expert manual annotation, we could know the news, ”Avril died in 2003”, is fake. With these historical statement labels, algorithms could learn patterns of reliable and unreliable data, and then predict the reliability of unknown data. On the other hand, in some situations, historical statement labels are not provided. Then, we have to look for other evidences to help differentiate reliable data and unreliable data.

In the meantime, values provided to one statement could be consistent or conflicting. Take an example with conflicting values: several websites (sources) provide different departing times (conflicting values) for one flight. Conflicting values may have different types, i.e., numeric or categorical values. Take an example with consistent values: in a social network, such as Twitter, people (sources) forward the news articles that they believe are real, and we could say that the ”Action: Forward” are consistent values. We could regard consistent values as the votes from providers to statements. Therefore, a robust framework that can handle different types of values is needed.

Scenario 1: sources provide consistent values to statements, and historical statement labels are not provided.

In this scenario, we could only observe data sources providing consistent values for statements, and the values are more like ”votes.” Without historical statement labels differentiating reliable and unreliable data, we have to collect other kinds of evidence. We could assume that the popularity indicates the reliability (unreliability, if votes mean negative support). Sometimes, the assumption may not hold real-world problems, e.g., both fake news and real news may have a large quantity of ”forwards” on social platforms. Without reading the news content or labeled training data, it is unreasonable to decide news veracity based on its popularity. Therefore, in Scenario 1, we need to collect reasonable background knowledge that could help differentiate reliable data from unreliable data.

In this study, we will work on a real-world problem, finding the cancer driver gene, to explore this scenario. Gene mutation randomly happens in chromosomes and could accumulate when people get older. Some gene mutations are cancer driver mutations that lead to cancer, while other mutations are passenger mutations, which has nothing to do with cancer. If a gene frequently mutates in cancer patients' cell samples, especially among those who have very few mutated genes, it is very likely to be a cancer driver gene.

The Cancer Genome Atlas (TCGA) is a landmark cancer genomics program, held by National Cancer Institute and the National Human Genome Research Institute, recording the patient gene mutation data over 33 cancer types (http://cancergenome.nih.gov/). We will identify cancer driver genes based on the cancer patients' gene mutation distribution in TCGA. To be more specific, data sources are patients; for each gene, the corresponding statement is "gene ** is a cancer driver gene"; if a gene mutates in a patient cell samples, we say the patient provides a value "yes" to the corresponding statement. We illustrate such a problem in Figure 3.

After data reliability assessed, we could rank statements by their reliability scores, but deciding statement veracity could be challenging. Without labeled training data, we will have to pick a heuristic strategy, such as selecting top N, or decide a threshold.

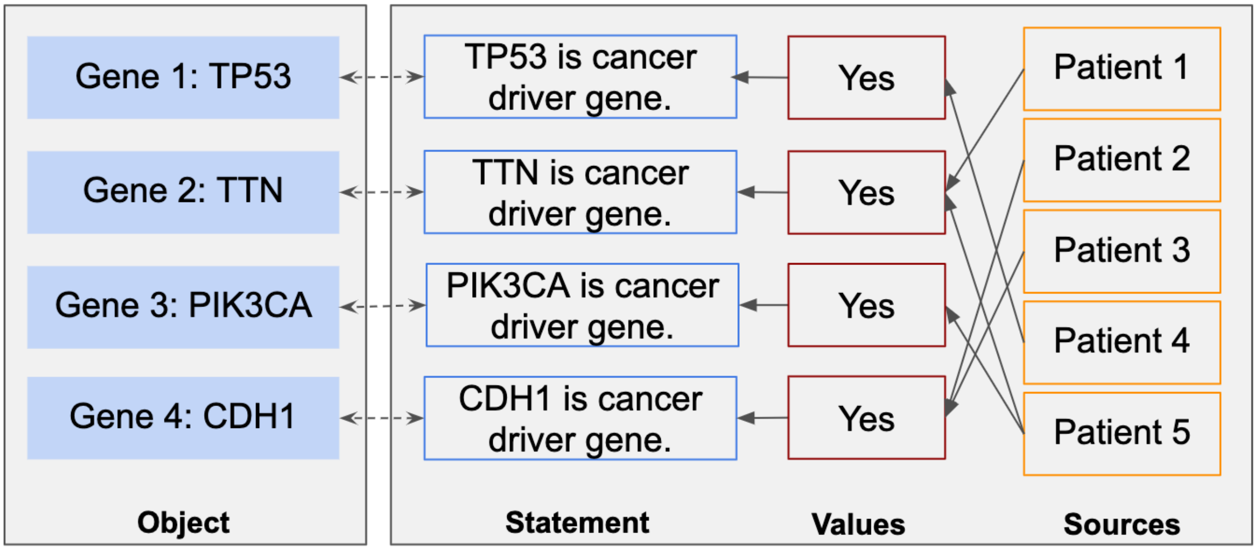


Figure 3: An example of sources providing consistent values. Cancer patients have mutation on both cancer driver genes and not-related genes.

Scenario 2: sources provide consistent values to statements, and historical statement labels are provided.

In this scenario, with historical statement labels, we could learn the characteristics and patterns of reliable sources and unreliable sources. Then, based on source reliability, we could predict the veracity of unlabeled statements.

Fake news propagation is an example of such a scenario. On social platforms, users forward the news. Some users can identify fake news and do not forward them, while other users are confused and forward many pieces of fake news. Please note when a user does not forward the news, maybe he believes the news is fake, or maybe they think the news is not interesting, or the news is not fed to his personalized account. Due to this reason, in this study, we only consider ”forward” as a positive vote, and do not take ”no forward” as a negative vote. This is why we do not consider the values as conflicting values. In such a case, the statement about a piece of news could be ”this news is true,” data sources are social platform users, and when they share the news, they provide value to the statement that ”Yes, I agree that news is true.” Such an example scenario is illustrated in Figure 4.

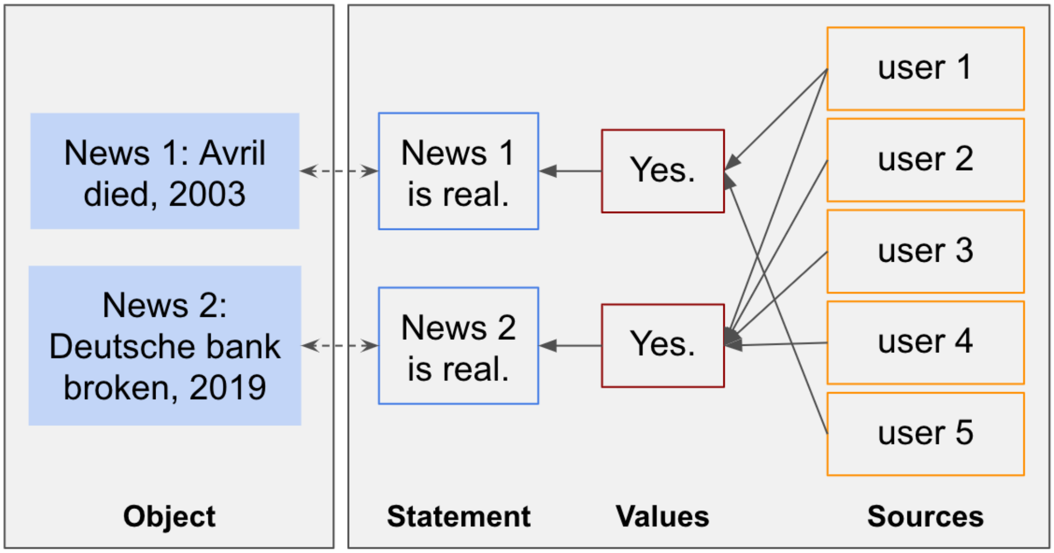


Figure 4: An example of sources providing consistent values. In social media platforms, users forward the news that they believe is real.

Scenario 3: sources provide conflicting values to statements, and historical statement labels are not provided.

In this scenario, given one statement, the values from different sources may be conflicting. The goal of data reliability assessment is to figure out the most likely value for the statement, and then the false statement could be fixed with the predicted true value. Without labeled historical data, we usually assume that majority of sources are reliable, and the value popularity implicates its truthfulness. Then the value receiving most support could be selected as the true value.

An example is the conflicting book author lists in online bookstores. As shown in Figure 5, five bookstores provide different author lists for two books in their web-pages. The goal is to accurately assess the reliability of the source and candidate values of each statement and figure out the correct statements. Such kind of problems appears in a research area called Truth Discovery.

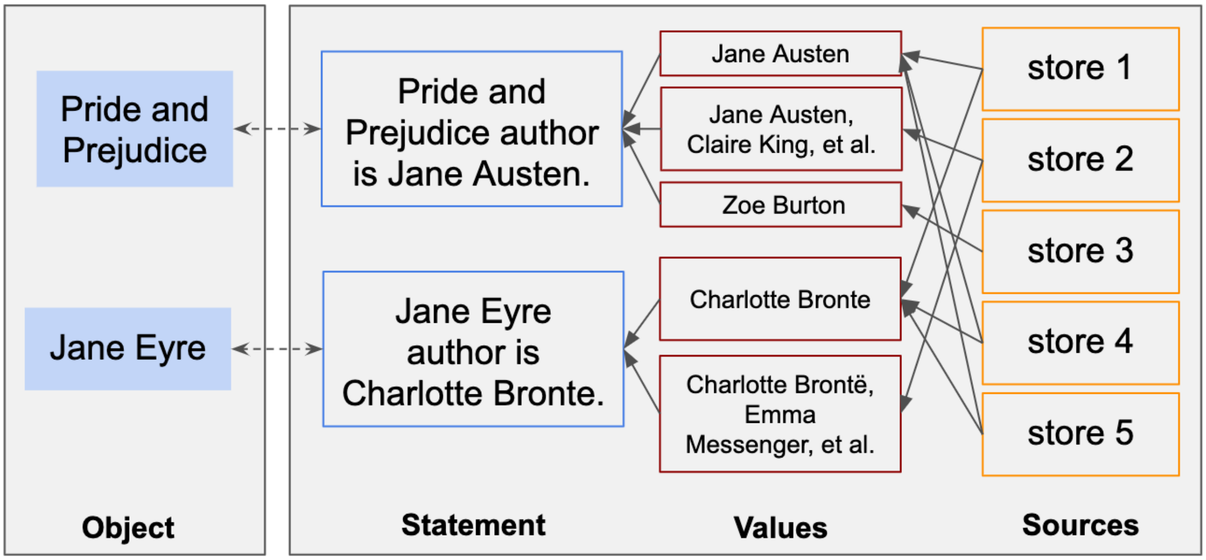

Pride and Prejudice
Jane Eyre
Object
Pride and Prejudice author is Jane Austen.
Jane Eyre author is Charlotte Bronte.
Statement
Jane Austen
Jane Austen, Claire King, et al.
Zoe Burton
Charlotte Bronte
Charlotte Brontë, Emma Messenger, et al.
Values
store 1
store 2
store 3
store 4
store 5
Sources

Figure 5: An example of sources providing conflicting values. Online bookstores provide different author list for two books.

Scenario 4: sources provide conflicting values to statements, and historical statement labels are provided.

In this scenario, with historical statement labels, we should explore patterns of reliable sources. Then later, based on assessed source reliability, we could predict the veracity of unlabeled statements.

An example is the editing of city populations in Wikipedia. As we know, any people can edit the wiki page (with modification approval), and there could be many different population data provided to the same city. As shown in Figure 6, five editors provide city populations for Pittsburgh and New York City, and we need to evaluate the editors' and population numbers' reliability and figure out the real population for each city. The goal is to accurately assess the reliability of the source and candidate values of each statement. A subset of city populations are known, and we could predict the rest based on them.

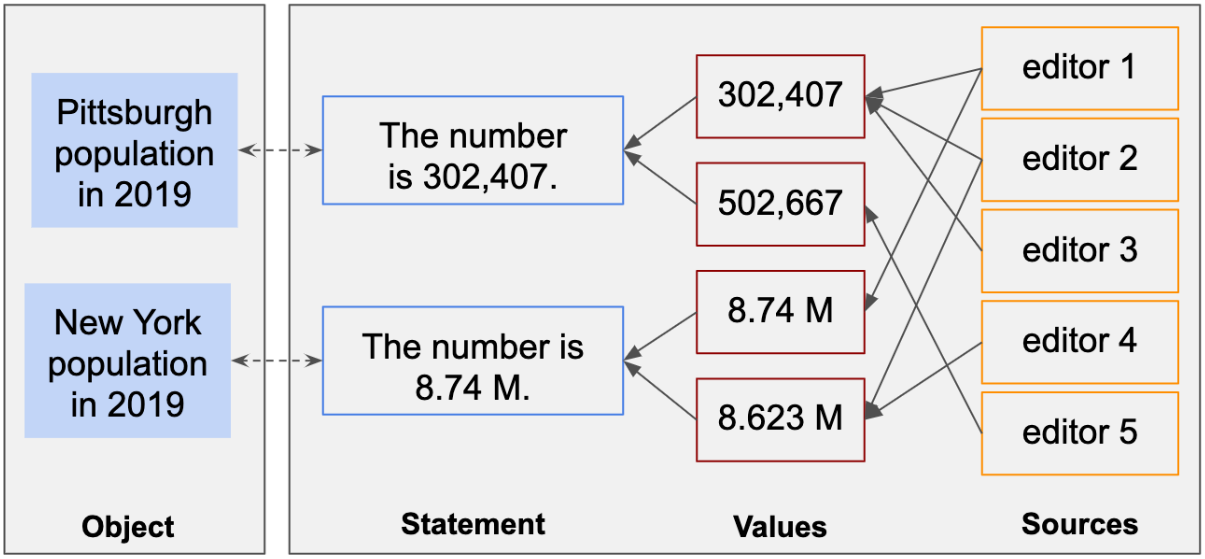


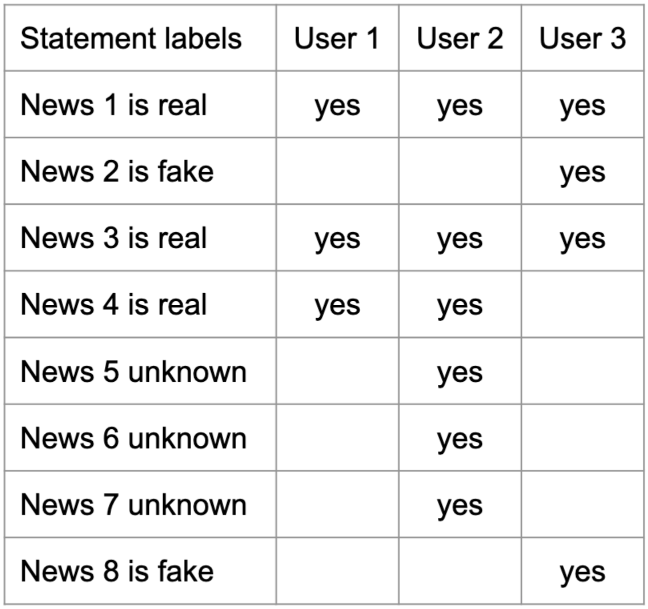
Figure 6: An example of sources providing conflicting values. Wikipedia editors provide different population data for two cities.

1.1.2 Subjective Opinion based data reliability representations

Data reliability assessment based on observed data without fully considering data uncertainty.

When we measure data reliability, it is natural to use the probability to represent it, such as the algorithm Accu from [Dong et al., 2009] . I.e., statement reliability could be defined as the probability of being real; source reliability could be defined as the probability of providing true values; given multiple conflicting values, value reliability could be defined as the probability of being true. For example, if a source always provides true values, then its reliability could be 100%.

In the area of fake news detection, truth discovery, and cancer driver gene discovery, the majority of past studies are using probability based models, maximizing the existence probability of the given dataset. However, these probability based representation effectiveness depends on the observed samples, and the possible uncertainty of the data is not taken into account, which will be illustrated by the following two examples.



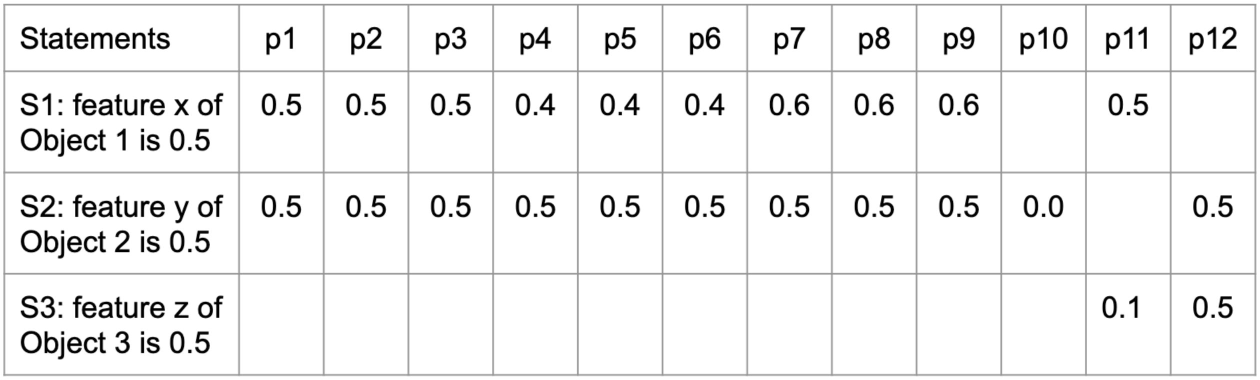
Statement labels	User 1	User 2	User 3
News 1 is real	yes	yes	yes
News 2 is fake			yes
News 3 is real	yes	yes	yes
News 4 is real	yes	yes	
News 5 unknown		yes	
News 6 unknown		yes	
News 7 unknown		yes	
News 8 is fake			yes

Figure 7: An example of data with uncertainty in fake news detection task.

Example 1. Figure 7 describes a simulated dataset in Scenario 2, where users share real/fake news on social platforms, and we have a part of historical statement labels. User 1 is 100% reliable user, as he spreads only real news; User 3 is 50% reliable user, as he spreads half real and half fake news; User 2 share three real news, and also share three other news whose labels are unknown yet. This unknown news veracity leads to uncertainty in the current data. For example, if news 5, 6, and 7 are all fake, then User 2 is as unreliable as User 3; if they are all real, then User 2 is as reliable as User 1.

If we use probability to measure reliability on labeled data, then the probability of User 2 sharing real news is 100% (3/3), same as User 1; if we use weight to measure reliability on labeled and unlabeled data, then User 2 weight could be 0.5 (3/6), same as User 3. Compared with User 1 and 3, User 2 has a different situation. As the situation uncertainty is not recorded, these reliability representations could not represent the whole information.

Example 2. Figure 8 describes a simulated dataset in Scenario 3, where sources (providers) provide conflicting values to statements, and no historical statement labels are provided. We could assume most providers have no malicious purpose, trying their best to provide true values. Among the conflicting values, mean value, median value, and the most popular value is very likely to be true.



Statements	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12
S1: feature x of Object 1 is 0.5	0.5	0.5	0.5	0.4	0.4	0.4	0.6	0.6	0.6		0.5	
S2: feature y of Object 2 is 0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.0		0.5
S3: feature z of Object 3 is 0.5											0.1	0.5

Figure 8: An example of data with uncertainty in truth discovery task.

In this simulated example dataset, common sense tells us that the true value of Statement 2 is very likely to be 0.5, as ten sources vote 0.5, and the other candidate value 0.0 seems to be an outlier. In terms of Statement 1, true value also has a chance to be 0.5, as four sources vote for 0.5, and the average of the other six values is also 0.5. In such a case, provider p11 provides ”true value” in Statement 1, and provider p12 provides ”true value” in Statement 2. Then the traditional probability or weight based methods would give high and same reliability scores for them. Then, it is hard to decide the true value for Statement 3, as two very reliable sources provide different values.

However, the situation of Statement 1 is more uncertain than that of Statement 2, as 0.4 and 0.6 have three votes respectively and have a chance of being true. The value variance in Statement 1 is big, so p11 has a higher probability that it did not provide true value. The above calculation is based on the assumption that the popularity wins and mainstream sources win. In the real world, it does happen that minority wins. So, we should have a high uncertainty opinion towards p11. On the other hand, p12 is very likely to provide true value. If we could consider such uncertainty, then in terms of Statement 3, maybe 0.5 from p12 is more likely to be true than 0.1 from p11.

In summary, traditional probability and weight based models lose the information about data uncertainty, and in turn, fail to characterise the data completely. In this dissertation study, we aim to find a better representation to assess data reliability in different scenarios.

Subjective Opinions based data reliability assessment.

Subjective Opinions could represent the probability affected by the degrees of uncertainty, i.e., express a person's subjective belief about the truth of a statement with degrees of uncertainty [Jøsang, 2016b]. To handle Subjective Opinions, Subjective Logic (SL) [Jøsang, 2016c] is used in this dissertation. A subjective opinion from a source p towards a statement s can be represented by a triple $\omega_s^p = \{t, d, u\}$, with $t, d, u \in [0, 1]$, and $t + d + u = 1$, where t means trust, d means distrust, and u means uncertainty. Traditional probability could be obtained as $(t+u/2)$. When $u = 0$, t equals probability, indicating that given no uncertainty, subjective opinion representation is same as in the probabilistic logic.

Then, when we decide about the statement veracity and select the most likely true values, we could focus on trust t and distrust d. They represent our opinions with pure certainty. On the other hand, the probability and weight based representations mix the certainty with uncertainty. Thus, Subjective Opinion based decision making should be more accurate.

In Example 1, User 1 only shares real news, and our subjective opinion could be $\omega^I_{User\ 1\ is\ reliable} = \{1, 0, 0\}$; User 3 shares half real and half fake news, and our subjective opinion could be $\omega^I_{User\ 3\ is\ reliable} = \{0.5, 0.5, 0\}$; User 2 shares 3 real news, and 3 unknown news, our subjective opinion could be $\omega^I_{User\ 2\ is\ reliable} = \{0.5, 0, 0.5\}$. In this way, we could easily observe the difference among the three users.

In Example 2, Statement 2 is relatively more certain than Statement 1. Though p11 and p12 give "true value" respectively, I have different subjective opinions towards their reliability, such as $\omega^I_{p11\ is\ reliable} = \{0.4, 0, 0.6\}$ and $\omega^I_{p12\ is\ reliable} = \{0.91, 0, 0.09\}$. Then in terms of Statement 3, since we trust more on p12, we may prefer 0.5 as the true value.

In this dissertation, we propose to use Subjective Opinion to represent our assessment for data reliability and expect this representation could have a better performance than traditional probability and weight based representations. Therefore, we expect to predict statement veracity and select true value more accurately.

SL is a calculus for Subjective Opinions, enabling us to manipulate the generated subjective opinions [Jøsang, 2016c]. Two Subjective Logic operations are utilized in this study. We have a subjective opinion towards a source, and the source recommends us with a value with some degree of uncertainty, then ”recommendation operator” in Subjective Logic could help us obtain our opinion towards the value. Also, when several sources provide value, we need to fuse these many subjective opinions, and the Subjective Logic ”consensus operator” could help combine them. They will be explained in Chapter 2.

1.2 Problem representation

Consider a dataset that contains a set of statements $S = \{s_1, s_2, ..., s_i, ..., s_n\}$, and a set of providers (sources) $P = \{p_1, p_2, ..., p_j, ..., p_m\}$. Each statement corresponds to an object, which is not explicitly shown in the data. Data sources provide values to the statements, and the value from a provider p_j to a statement s_i is defined as v_{ij}. We have the following mapping function:

$$f : P \times S \to V. \tag{1}$$

Please note that $v_{ij} = null$ means provider p_j doesn't provide values to statement s_i. Such a dataset can be represented as a matrix, as shown in Table 2. The providers' and the values' reliability needs to be assessed, and the statements' veracity needs to be predicted.

In terms of statements and providers, we have the following assumptions:

- **data sources are independent of each other**, i.e., while providing the values, a data source does not reference or copy other sources' values.
- **statements are independent of each other**, i.e., statements relationship, such as similarity, conflict, and relatedness, are not considered.

- **majority data sources have no malicious purpose**, i.e., providing false values deliberately.

In this dissertation, I explore how to map different scenarios into this problem representation and how to more accurately assess the data reliability with Subjective Opinions handling data uncertainties.

Table 2: The dataset is represented as a matrix, with n statements and m sources/providers.

	$\mathbf{p_1}$	$\mathbf{p_2}$	$\mathbf{p_3}$	**…**	$\mathbf{p_j}$	**…**	$\mathbf{p_m}$
$\mathbf{s_1}$	v_{11}	v_{12}	v_{13}	…	v_{1j}	…	v_{1m}
$\mathbf{s_2}$	v_{21}	v_{22}	v_{23}	…	v_{2j}	…	v_{2m}
$\mathbf{s_3}$	v_{31}	v_{32}	v_{33}	…	v_{3j}	…	v_{3m}
…	…	…	…	…	…	…	…
$\mathbf{s_i}$	v_{i1}	v_{i2}	v_{i3}	…	v_{ij}	…	v_{im}
…	…	…	…	…	…	…	…
$\mathbf{s_n}$	v_{n1}	v_{n2}	v_{n3}	…	v_{nj}	…	v_{nm}

1.3 Research questions

Research Question 1: How to assess data reliability when sources provide consistent values to statements without historical data?

Research Question 2: How to assess data reliability when sources provide consistent values to statements, and historical data are provided?

Research Question 3: How to assess data reliability when sources provide conflicting values to statements, and historical data are not provided?

Research Question 4: How to assess data reliability when sources provide conflicting values to statements, and historical data are provided?

For each of the above research questions, we explore the following questions:

- What are the uncertainties in this data scenario?
- How to comprehensively record the uncertainty in the reliability representation?
- How to assess data reliability with such representation?
- Is such data reliability assessment more accurate than past models'?

1.4 Overview of the Chapter Structure

The dissertation includes seven chapters. Chapter 1 introduces the background and motivation of this dissertation study, and also specific the four research questions. Chapter 2 introduces the related works. Chapter 3 to Chapter 6 details the proposed solutions for each research question. Chapter 7 provides the conclusion.

2.0 Related Works

In this chapter, we will first explore the area of data reliability assessment area, introduce its history and recent works. Then, based on the real-world problems we will work on, background knowledge and related works of three areas will be discussed, including fake news detection, truth discovery, and cancer driver gene discovery. Also, as our data reliability assessment is based on Subjective Opinions, the related works and more background knowledge of it and Subjective Logic are considered.

2.1 Data Reliability Assessment

Data reliability has always been considered as a very important issue. Data come from diverse sources, fluctuating our daily lives. People make decision based on the information obtained from such data. However, data quality usually cannot be guaranteed, and hence decision making with such unreliable data may lead to loss beyond estimation. For example, in the 2016 US presidential election season, fake news about Hillary Clinton fluctuated social media and had an indispensable effect on the election results. Accurate data reliability assessment mechanisms can help people or machines recognize the distrustful information and then filter unreliable data out. Afterward, people can make decisions with clean, consistent, and reliable information.

Barlow et al., defines reliability as a quantified measure of uncertainty about a particular type of event [Barlow and Proschan, 1975]. For example, assume an event is ”source A provide true values”, the reliability could describe our measure towards this event's veracity. Probability is a method to quantify the reliability [Bennett et al., 2003], and in this dissertation, we propose to assess reliability with Subjective Opinions.

In the area of integrated databases, Zadorozhny and Grant [2016] classify the reliability into internal reliability and external reliability. Internal reliability is about the value inconsistency without considering data sources. For example, in a population dataset, New York

City's population is bigger than New York State's population, which is obviously inconsistent. There are many studies working on the internal reliability assessment and corresponding data inconsistency resolution, such as methods in [Dong and Naumann, 2009, Bleiholder and Naumann, 2009, Bertossi, 2006, Bertossi and Chomicki, 2004, Destercke et al., 2011].

By the definition from [Zadorozhny and Grant, 2016], the reliability explored in this dissertation is the external reliability, which is outside of the database and take the reliability of the data sources into consideration. Such kind of external data reliability assessment plays an important role in the Truth Discovery and Fake News Detection, and will be discussed more in the following two sections.

Wireless Sensor Network (WSN) is another type of data structure and reliability assessment is an important procedure to get good quality data [Jaggle et al., 2009]. In a WSN, many sensor nodes collect data from the environment, and there is a reader could read data from sensors. However, the communication of sensor-sensor and the sensor-reader may fail, and the collected data reliability is not guaranteed. Hence, data reliability assessment is a challenging topic in such a scenario. There are many models proposed, such as [Jaggle et al., 2009, Chiu et al., 2001, Hardy et al., 2007, Shpungin, 2007].

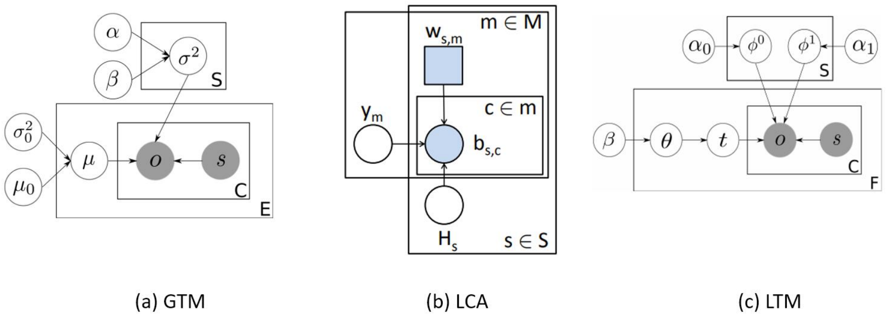
Additionally, data reliability assessment are explored and discussed from other aspects. Quinn et al. [2009] proposed a framework that allows user community to annotate their trust towards the data and hence reliability is collected. Götzinger et al. [2017] deployed a hierarchical agent-based system that classifies data reliability but using Fuzzy logic instead of conventional Boolean values, and use it to enhance the Early Warning Score (EWS) systems in health-care domain. Shcherbakov et al. [2005] proposed a method of employing the principal component analysis (PCA) to assess cirrus cloud data reliability.

2.2 Truth discovery

In this section, we introduce the related works in Truth Discovery area. It mainly solves the data conflict problem in data fusion by selecting the true value from multiple candidate values for each statement, and accurate data reliability assessment is important [Li et al., 2016]. After reviewing the existing literature, I classify them into two categories: (1) single true value truth discovery methods, (2) multiple true values truth discovery methods. In this study, we focus on the first category with ”single true value”, which can be further separated into five groups: (1) Bayesian based models, (2) web links analysis based models, (3) information retrieval based methods, (4) crowd sourcing based models, (5) external knowledge based models.

The first group is the early Bayesian based methods, which in the procedure usually generates a temporary true value used for parameter calculation. Dong et al. [2009], proposed a method called Accuracy, which is calculated as the probability of each value being correct, and average the confidence of facets provided by the source as the provider trustworthiness. After that, they proposed the concept of Accuracy-similarity, which further considers the similarity of two values. Then, Dong et al. [2012] proposed POPAccuarcy, which differs from Accuracy by releasing the assumption that false value probability is uniformly distributed. In [Dong et al., 2009, 2012] researchers also explored the data copying problem. Another Bayesian method is the TruthFinder, proposed by Yin et al. [2008], which differs from Accuracy by not normalizing the confidence score of each statement. Dong et al. [2015] further proposed to learn the web source trustworthiness through a multi layer probabilistic model, where they assume many extractors work on extracting facts from online web pages, and the source in different extractors, should have a different trustworthiness.

The second group of methods is also Bayesian based methods, and more particularly, can be called as probabilistic graphical model based methods. Zhao and Han [2012] proposed Gaussian Truth Model (GTM), in which the data source reliability follows a Gaussian distribution, and value reliability follows a gamma distribution, as shown in Figure 9 (a). The parameters are learned through EM. Pasternack and Roth [2013] proposed a similar model, Latent credibility analysis (LCA), and additionally allows the data source to have a confidence score on the data it provides, as shown in Figure 9 (b). Zhao et al. [2012] proposed a similar graphical model, but define the reliability of data source by both sensitivity and specificity, which evaluate the data source reliability on both true positive and true negative values, as shown in Figure 9 (c). Also, it differentiates the data source’s value of agree (provide the value), disagree (provide other value), and null.



(a) GTM (b) LCA (c) LTM

Figure 9: Different probabilistic graphical models in truth discovery.

This third group of methods are regarded as the optimization problem, which aims to find the most appropriate trustworthiness score and value truth assignment to get optimization under some constraints. Li et al. [2014a] proposed a confidence aware approach for truth discovery (CATD) to minimize the distances between true value and fake values, and also minimize the weight of data sources generating big errors. Li et al. [2014b] generalize CATD by allowing the value to be heterogeneous, and distances between true value and fake values can be calculated and aggregated from different dimensions. Rekatsinas et al. [2017] proposed SLiMFast where data source weight is tuned to maximum the probability of the observed data.

The forth group of methods are based on the web links analysis. Pasternack and Roth [2010] proposed three methods: (1) AverageLog is a transformation of Hub-Authority algorithm, with source trustworthiness being the averaged confidence score of provided values multiplying the log of provided value count; (2) Investment, where the confidence score of the value grows exponentially with the accumulated providers' trustworthiness. (3) Pooled-Investment, where the confidence score of the value grows linearly. Galland et al. [2010] proposes 2-Estimates, which is a transformation of Hub-Authority algorithm, whose provider trustworthiness is the average instead of the sum of the vote count. They further proposed 3-Estimates, which additionally considers the difficulty to get the true values.

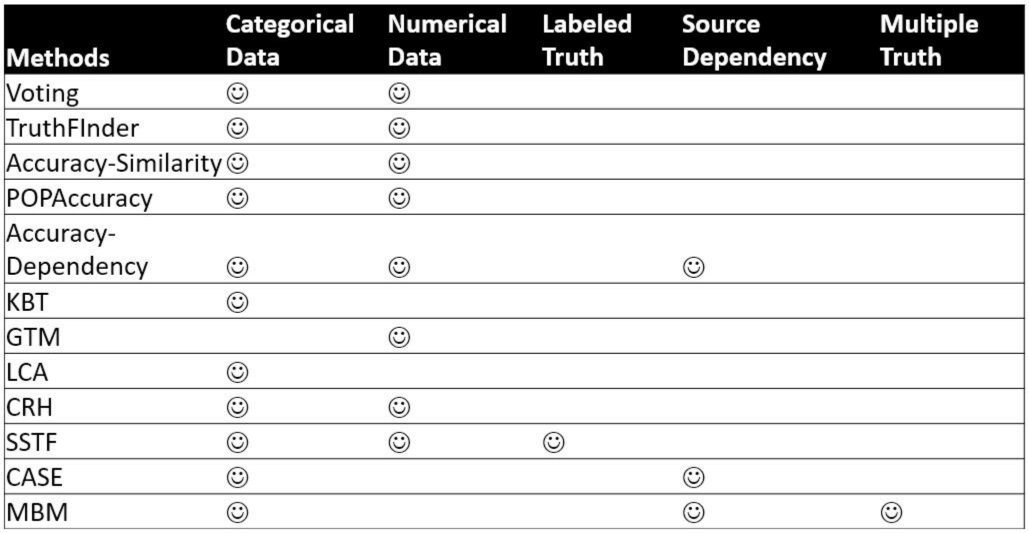
The fifth group of methods are semi-supervised methods, which will utilize a subset of data with groundtruth. Yin and Tan [2011] proposed a semi-supervised reliability assessment method, SSTF. It is basically a PageRank method assuming that there is a set of statements having the true value, which will affect the result in the PageRank iteration. Mukherjee et al. [2016] proposed TruthCore that use a small set of training data to construct an independent reliable sources, starting from which the outliers are removed in each iteration to balance the similarity among the sources. Pochampally et al. [2014] proposed a method to measure the source precision and recall, and correlation (dependency) between sources, based on which the value confidence score is computed. They used extra training data to calculate the precision and recall.

Also, different from above mentioned method, a network embedding based method, CASE, is proposed [Lyu et al., 2017]. It constructs a network by data sources and values they provide, and embed the network into a low dimensional vector space. The true value will be the one who is most close to majority voting selected averaged embedding vector.

In addition, in recent years, with the increasing popularity of online crowdsourcing platform, truth discovery models are frequently used in this scenario to help pick the useful information [Ouyang et al., 2016, Nguyen et al., 2017, Zheng et al., 2016, Chen et al., 2017, Yin et al., 2017]. For example, people from crowdsourcing platforms provide different labels for arguments, Nguyen et al. [2017] proposed a method to aggregate the arguments. Ouyang et al. [2016] proposed a scalable and effective way to deal with the data fusion of the crowdsourcing data in a large scale.

As mentioned above, there is a set of models working on multiple-truth-discovery problems. Wang et al. [2015] proposed an integrated Bayesian approach to the multi-truth-finding problem. It consider the involvement of sets of values in claims, different implications of inter-value mutual exclusion, and larger source profiles. Then, Wang et al. [2016] further improved their method by taking the value implications into consideration (i.e., the data similarity), and also replace the Bayesian model with a probabilistic graphical model, as the work in [Zhao and Han, 2012]. Afterwards, Fang et al. [2017a,b] proposed a graph-based method, SmartMTD, models and quantifies two types of source relations to estimate source reliability precisely and to detect malicious agreement among sources for effective multi-truth discovery. We do not go deeper, as we only focus on single truth value problems.

Finally, we select 12 major methods, and compare them in Figure 10. From [Waguih and Berti-Equille, 2014, Fang et al., 2017c, Li et al., 2012] we can know that, there is no ”one-fits-all” approach that always beats others in all dataset. Further, according to [Waguih and Berti-Equille, 2014, Fang et al., 2017c], groundtruth data of current datasets that are used to evaluate and compare the methods are too sparse to generate statistically significant results. The groundtruth in popular datasets, such as Book-Author, Flight, Movie, and Biography, covers less than 10% objects in the dataset. In some scenarios, the groundtruth data is even unfeasible. Therefore, when faced with a real scenario, it maybe safe to try several methods, and construct a large enough validation dataset to select the fitting model(s), or try to compare the methods through the way like [Fang et al., 2017c] without groundtruth data.



Methods	Categorical Data	Numerical Data	Labeled Truth	Source Dependency	Multiple Truth
Voting	☺	☺			
TruthFInder	☺	☺			
Accuracy-Similarity	☺	☺			
POPAccuracy	☺	☺			
Accuracy-Dependency	☺	☺		☺	
KBT	☺				
GTM		☺			
LCA	☺				
CRH	☺	☺			
SSTF	☺	☺	☺		
CASE	☺			☺	
MBM	☺			☺	☺

Figure 10: An summary of 12 major methods in truth discovery area.

2.3 Fake news detection

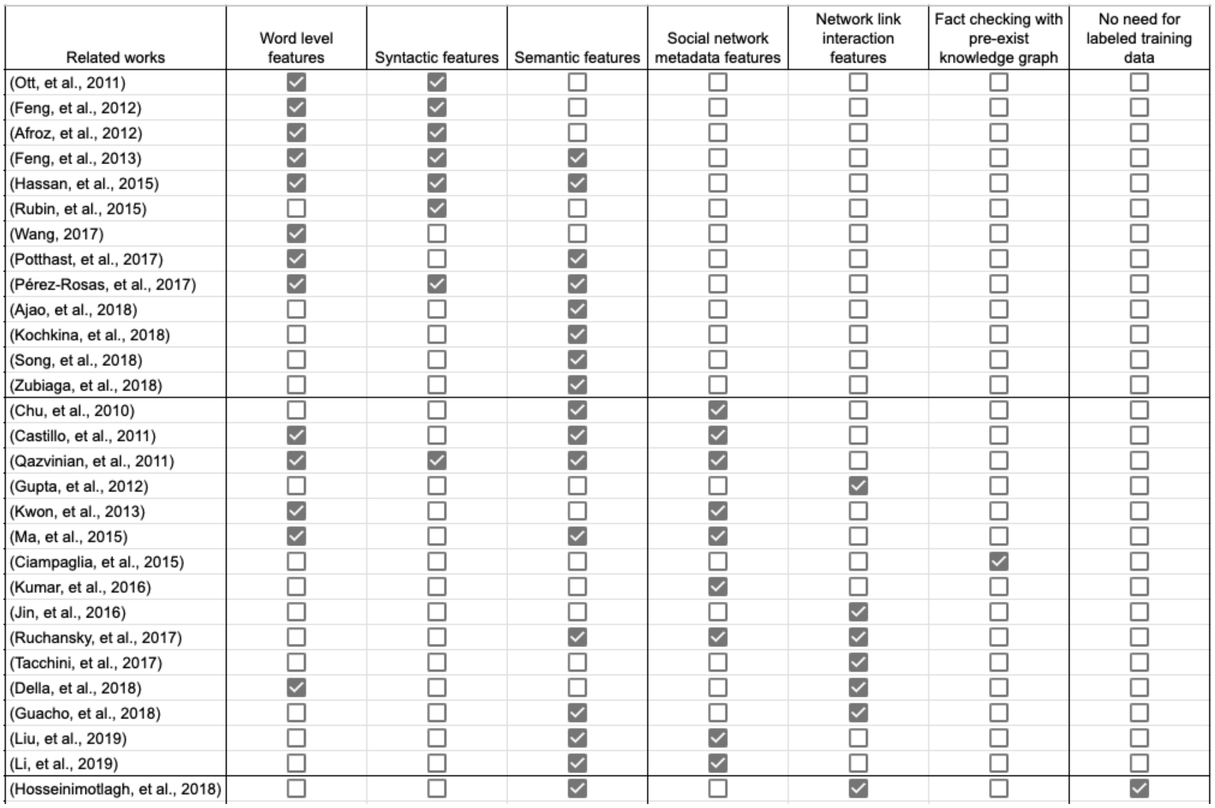
Several recent survey papers encompass the wide range of research devoted to fake news including [Kumar and Shah, 2018, Shu et al., 2017a, Conroy et al., 2015, Chen et al., 2015, Bondielli and Marcelloni, 2019]. The most important problem in this area is to automatic detection. As mentioned in the previous section, there are different types of fake news and there is also a close connection with rumours. So fake news detection techniques have a substantial overlap with the detection of rumours, fake opinion, fake accounts, hoaxes, and frauds. For that reason we include some algorithms from papers about those that can also be used for fake news detection.

Fake news detection mainly use three kinds of information: (1) the content of news articles, including word level, syntactic level, and semantic level information, (2) news propagation by users on social networks, including user profiles, news profiles, spreading data, etc, and (3) network structure extracted from news articles and social media. In most works, detection is implemented by a classification model on different kinds of features.

Word level and syntactic level features of news articles are found to be the most effective in many papers, such as models in [Rubin and Lukoianova, 2015] and [Wang, 2017]. Both word and syntactic information are essential. The increasing popularity of neural networks in natural language processing (NLP) has led to the extraction and use of semantic features in fake news detection, such as the models in [Hassan et al., 2015, Potthast et al., 2017, Pérez-Rosas et al., 2017, Ajao et al., 2018, Kochkina et al., 2018, Song et al., 2019, Zubiaga et al., 2018].

From another aspect, many researchers explore the information from the social network where the news is spread and to the people in the network. They focus on the news profile features, such as the number of likes and propagation times, and user profile features, such as the number of posts, registration age, and the number of followers. Many studies have found that systems cannot detect fake news accurately if they only use social network features, so they are usually used together with the content features, such as models in [Castillo et al., 2011, Chu et al., 2010, Qazvinian et al., 2011, Kwon et al., 2013, Ma et al., 2015, Kumar et al., 2016, Liu et al., 2019, Li et al., 2019].

Multiple network structures can be obtained from this area, such as user-follow-user networks, news-agree/conflict-news networks, and user-spread-news networks [Gupta et al., 2012, Jin et al., 2016, Ruchansky et al., 2017, Tacchini et al., 2017, Della Vedova et al., 2018, Guacho et al., 2018]. There is also a smaller number of works that focus on news fact checking, where the reference facts are in a preexisting knowledge base such as DBpedia [Wu et al., 2014, Ciampaglia et al., 2015, Shi and Weninger, 2016]. A comparison of related fake news detection studies is given in Figure 11.



Related works	Word level features	Syntactic features	Semantic features	Social network metadata features	Network link interaction features	Fact checking with pre-exist knowledge graph	No need for labeled training data
(Ott, et al., 2011)	☑	☑	□	□	□	□	□
(Feng, et al., 2012)	☑	☑	□	□	□	□	□
(Afroz, et al., 2012)	☑	☑	□	□	□	□	□
(Feng, et al., 2013)	☑	☑	☑	□	□	□	□
(Hassan, et al., 2015)	☑	☑	☑	□	□	□	□
(Rubin, et al., 2015)	□	☑	□	□	□	□	□
(Wang, 2017)	☑	□	□	□	□	□	□
(Potthast, et al., 2017)	☑	□	☑	□	□	□	□
(Pérez-Rosas, et al., 2017)	☑	☑	☑	□	□	□	□
(Ajao, et al., 2018)	□	□	☑	□	□	□	□
(Kochkina, et al., 2018)	□	□	☑	□	□	□	□
(Song, et al., 2018)	□	□	☑	□	□	□	□
(Zubiaga, et al., 2018)	□	□	☑	□	□	□	□
(Chu, et al., 2010)	□	□	☑	☑	□	□	□
(Castillo, et al., 2011)	☑	□	☑	☑	□	□	□
(Qazvinian, et al., 2011)	☑	☑	☑	☑	□	□	□
(Gupta, et al., 2012)	□	□	□	□	☑	□	□
(Kwon, et al., 2013)	☑	□	□	☑	□	□	□
(Ma, et al., 2015)	☑	□	☑	☑	□	□	□
(Ciampaglia, et al., 2015)	□	□	□	□	□	☑	□
(Kumar, et al., 2016)	□	□	□	☑	□	□	□
(Jin, et al., 2016)	□	□	□	□	☑	□	□
(Ruchansky, et al., 2017)	□	□	☑	☑	☑	□	□
(Tacchini, et al., 2017)	□	□	□	□	☑	□	□
(Della, et al., 2018)	☑	□	□	□	☑	□	□
(Guacho, et al., 2018)	□	□	☑	□	☑	□	□
(Liu, et al., 2019)	□	□	☑	☑	□	□	□
(Li, et al., 2019)	□	□	☑	☑	□	□	□
(Hosseinimotlagh, et al., 2018)	□	□	☑	□	☑	□	☑

Figure 11: An summary of 29 major methods in fake news detection area.

2.4 Cancer driver gene mutation discovery

In cancer genomics, one of the most critical tasks is to distinguish cancer driver mutations from passenger mutations [Greenman et al., 2007]. In human cells, gene mutates randomly. The abnormal behaviors demonstrated by cancer cells are the result of a series of mutations in key regulatory genes (https://www.cancerquest.org/), and we call these mutations as cancer driver mutations. The other mutations that do not lead to cancer cells are called passenger mutations. The most simple methods are based on mutation frequency, i.e., most frequently mutated genes in cancer patient cells are driver genes [Ding et al., 2008]. However, such a naive model classifies too many non-related genes as driver genes (high False Discovery Rate) [Banerji et al., 2012], and many advanced models are proposed.

The first type of model is based on the Background Mutation Rate (BMR). Genes have a "normal" mutation rate, which is called background mutation rate, and if they mutate more often than expected in cancer cells, the genes are likely to be cancer driver genes. Algorithms based on such evidences include MutSig [Banerji et al., 2012], MutSigCV [Lawrence et al., 2013] and MutSiC [Dees et al., 2012], Simon [Youn and Simon, 2011], OncodriverFM [Gonzalez-Perez and Lopez-Bigas, 2012] and ActiveDriver [Reimand and Bader, 2013]. Michaelson et al. [2012] finds that the gene BMR variety could be very large. Considering this, MutSigCV uses patient-specific mutation frequency and spectrum, and gene-specific BMR. Also, OncodriverFM uses external knowledge, the functional impacts of mutation, to help identify the driver genes. Our baselines will be selected from this type of model.

The second type of model is based on machine learning and labeled training data. CanPredict [Polymorphisms, 2007], CHASM [Carter et al., 2009] used Random Forest model to learn driver gene patterns. SVM is used in [Jordan and Radhakrishnan, 2014]. Census of Human Cancer Genes (CGC) provides a list of cancer genes, which is usually used as labeled driver genes. However, there is no golden standard passenger gene list. For example, in [Carter et al., 2009], synthetic passenger mutations are generated to play as labeled passenger genes; in [Tan et al., 2012], a list of none disease driver genes from (https://www.uniprot.org/docs/humsavar) are used as labeled passenger genes. As mentioned above, TCGA is a landmark cancer genomics program, held by the National Cancer

Institute and the National Human Genome Research Institute, recording the patient gene mutation data over 33 cancer types (http://cancergenome.nih.gov/). In this dissertation, our utility case with TCGA data considers the scenario without labeled training data, and hence this type of model is not selected as baselines.

The third type of model is the network and pathway-based approaches, which consider the gene-gene interaction. Such kind of models include MEMo [Ciriello et al., 2012], Dendrix [Vandin et al., 2012], PARADIGM-Shif [Ng et al., 2012], DriverNet [Bashashati et al., 2012], TieDIE [Paull et al., 2013], MAXDRIVER [Chen et al., 2013] and DawnRank [Hou and Ma, 2014]. In this dissertation, we only used the gene mutation data and did not consider the gene network and pathway knowledge, and hence this type of model is not selected as baselines.

Besides, there are methods using other medical knowledge and data to help identify driver genes. For example, CONEXIC introduces the gene copy number change into their model. In this dissertation, we will only use the gene mutation data from TCGA.

2.5 Subjective Opinions and Subjective Logic

Subjective Logic [Jøsang, 2016a, 1997] is a powerful decision making tool extending the probabilistic logic by including uncertainty and subjective belief ownership. It is widely used in trust network analysis [Jøsang et al., 2006], conditional inference [Josang, 2008], information provider reliability assessment [Pelechrinis et al., 2015], trust management in sensor networks [Oleshchuk and Zadorozhny, 2007], etc. Subjective logic uses subjective opinions to express subjective beliefs about the truth of propositions with degrees of uncertainty. Kane and Browne [2006] successfully applied subjective logic to a wireless network environment. Liu et al. [2011] presented a novel reputation computation model to discover and prevent selfish behaviors by combining familiarity values with subjective opinions. To the best of our knowledge, our work is the first one applying SL to area of data reliability assessment.

Subjective Logic defines a set of logical operations [Oleshchuk and Zadorozhny, 2007], and in this paper we use two of them:

- Recommendation operation. Assume two persons, A and B: A has an opinion towards B, and B has an opinion towards a statement s. Then according to B's recommendation, A also has an opinion towards this statement s. The definition of tje recommendation operator $\otimes$ is as follows:
  $\omega_S^{AB} = \omega_B^A \otimes \omega_S^B = \{t_S^{AB}, d_S^{AB}, u_S^{AB}\}$,
  where $t_S^{AB} = t_B^A t_S^B$, $d_S^{AB} = t_B^A d_S^B$,
  and $u_S^{AB} = d_B^A + u_B^A + t_B^A u_S^B$.
- Consensus operation. If two persons A and B have opinions towards one statement S, then consensus operator $\oplus$ can be used to combine their opinions. The definition of the consensus operator $\oplus$ is as follows:
  $\omega_S^{A,B} = \omega_S^A \oplus \omega_S^B = \{t_S^{A,B}, d_S^{A,B}, u_S^{A,B}\}$,
  where $t_S^{A,B} = (t_S^A u_S^B + t_S^B u_S^A)/(u_S^A + u_S^B - u_S^A u_S^B)$,
  $d_S^{A,B} = (d_S^A u_S^B + d_S^B u_S^A)/(u_S^A + u_S^B - u_S^A u_S^B)$,
  and $u_S^{A,B} = (u_S^A u_S^B)/(u_S^A + u_S^B - u_S^A u_S^B)$.

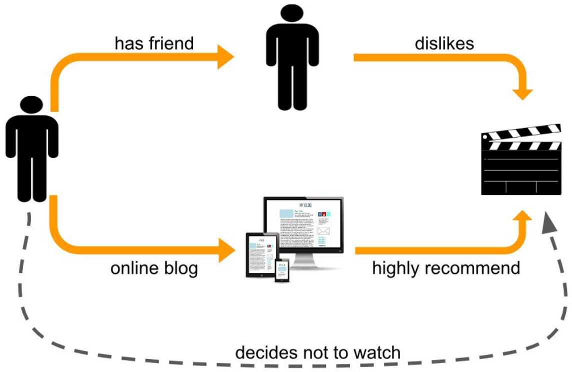


Figure 12: The procedure of deciding whether to watch a movie.

With the recommendation and consensus operations, people can merge their opinions towards an unknown statement. For example, person A wants to know whether a new movie is worthy to watch, as shown in Figure 12. He searches online and finds a blog saying the movie is absolutely the best movie of the year. However, his friend B told him that he watched the movie yesterday, and it is just an average work. We can assume A's opinion towards B (i.e., the statement "B is trustful") is {0.8, 0.1, 0.1}, while B's impression towards the movie (i.e., the statement "the movie worth my money and time") is {0.5, 0.4, 0.1}. Then, according to B's recommendation, A will have an opinion towards the movie $\{t^{AB}_{movie}, d^{AB}_{movie}, u^{AB}_{movie}\}$, where

$$t^{AB}_{movie} = t^A_B t^B_{movie} = 0.8 * 0.5 = 0.4,$$
$$d^{AB}_{movie} = t^A_B d^B_{movie} = 0.8 * 0.4 = 0.32,$$
$$u^{AB}_{movie} = d^A_B + u^A_B + t^A_B u^B_{movie} = 0.1 + 0.1 + 0.8 * 0.1 = 0.28.$$

It is possible that the blog is an advertisement to attract people with overpraised words, and we assume A's opinion about the blog is {0.5, 0.2, 0.3}. The blog holds an opinion {1, 0, 0} towards the movie. According to the blog's recommendation, A's opinion towards the movie is $\{t^{A,blog}_{movie}, d^{A,blog}_{movie}, u^{A,blog}_{movie}\}$, where

$$t^{A,blog}_{movie} = t^A_{blog} t^{blog}_{movie} = 0.5 * 1 = 0.5,$$
$$d^{A,blog}_{movie} = t^A_{blog} d^{blog}_{movie} = 0.5 * 0 = 0,$$
$$u^{A,blog}_{movie} = d^A_{blog} + u^A_{blog} + t^A_{blog} u^{blog}_{movie} = 0.2 + 0.3 + 0.5 * 0 = 0.5.$$

After combining both friend B and blog's recommendation via consensus operation, A has a final impression towards the movie, $\{t^{A,B;A,blog}_{movie}, d^{A,B;A,blog}_{movie}, u^{A,B;A,blog}_{movie}\}$, where

$t^{A,B;A,blog}_{movie} = (t^{A,B}_{movie} u^{A,blog}_{movie} + t^{A,blog}_{movie} u^{A,B}_{movie})/(u^{A,B}_{movie} + u^{A,blog}_{movie} - u^{A,B}_{movie} u^{A,blog}_{movie}) = (0.5 * 0.5 + 0.5 * 0.3)/(0.3 + 0.5 - 0.3 * 0.5) = 0.5313,$

$d^{A,B;A,blog}_{movie} = (d^{A,B}_{movie} u^{A,blog}_{movie} + d^{A,blog}_{movie} u^{A,B}_{movie})/(u^{A,B}_{movie} + u^{A,blog}_{movie} - u^{A,B}_{movie} u^{A,blog}_{movie}) = (0.2 * 0.5 + 0 * 0.3)/(0.3 + 0.5 - 0.3 * 0.5) = 0.25,$

$u^{A,B;A,blog}_{movie} = (u^{A,B}_{movie} u^{A,blog}_{movie})/(u^{A,B}_{movie} + u^{A,blog}_{movie} - u^{A,B}_{movie} u^{A,blog}_{movie}) = (0.3 * 0.5)/(0.3 + 0.5 - 0.3 * 0.5) = 0.2188.$

Finally, A decides to watch the movie.

3.0 RQ 1: how to assess data reliability when candidate values are not conflict and and historical data is not provided

Scenario 1 is the most challenging scenario in this study. There is no historical statement labels, and each statement only has one value from different sources. We could say sources are voting statements. In first section, we give our proposed Subjective Opinion based framework for data reliability assessment, and in second section, we will work with a real-world data reliability assessment problem, see how the framework is actually utilized and evaluate its performance.

3.1 Subjective Opinion based data reliability assessment

In my Subjective Opinion based data reliability assessment framework, statement reliability is represented by $\omega_{s_i}^{model}$, describing the model's opinion towards the declaration "statement s_i is real", and provider reliability is represented by $\omega_{p_j}^{model}$, describing the model's opinion towards the declaration "provider p_j is real". My framework first initialize the statement reliability, then iteratively assess provider/source reliability and statement reliability until converge, and lastly predict statement veracity based on reliability, as shown in Figure 13.

Step-1: initialize statement reliability. Initially, all the statements receive same reliability, $\{1, 0, 0\}$. It indicates that initially we believe all statements are true, and their reliability will be calculated in the iteration.

Step-2: evaluate provider reliability. Provider reliability is defined as model's opinion towards the statement "provider p_j is reliable", which is defined as:

$$\omega_{p_j}^{model} = \{t_{p_j}^{model}, d_{p_j}^{model}, u_{p_j}^{model}\}. \tag{2}$$

The task background knowledge will help define the specific values of the triple, and hence this part should be different in different tasks.

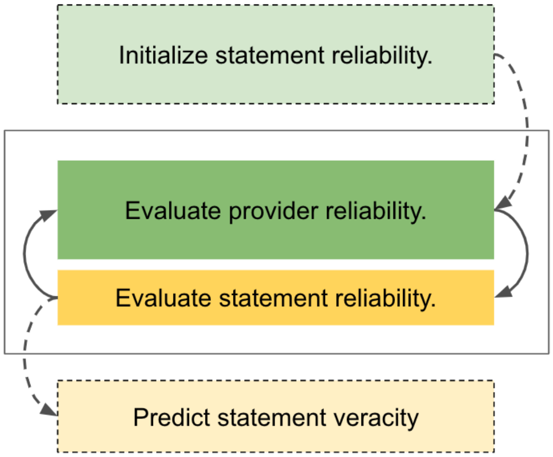

Initialize statement reliability.
Evaluate provider reliability.
Evaluate statement reliability.
Predict statement veracity

Figure 13: Subjective Opinion based data reliability assessment framework for Scenario 3.

Step-3: evaluate statement reliability. For each statement, we collect all sources who provide values to it. Based on whether provider has same belief in all statements that he provides values to, we have two different statement reliability assessment cases.

Case 1 If provider/source p_j has same belief in all statements that he provide values to, then directly fuse these sources' reliability with Subjective Logic consensus operation. The fused opinion represents our belief for the statement "statement s_i is real".

$$\omega_{s_i}^{model} = \omega_{p_1}^{model} \oplus \omega_{p_2}^{model} \oplus ... \oplus \omega_{p_k}^{model}, \tag{3}$$

where p_1, p_2,..., p_k are sources providing values to s_i.

Case 2 If provider/source p_j has different beliefs in the statements that he provide values to, first get the Subjective Opinion to statements according to provider's recommendation, and then fuse these opinions with Subjective Logic consensus operation. The fused opinion represents our belief for the statement "statement s_i is real".

$$\omega_{s_i}^{model} = (\omega_{p_1}^{model} \otimes \omega_{s_i}^{p_1}) \oplus (\omega_{p_2}^{model} \otimes \omega_{s_i}^{p_2}) \oplus ... \oplus (\omega_{p_k}^{model} \otimes \omega_{s_i}^{p_k}), \tag{4}$$

where p_1, p_2,..., p_k are sources providing values to s_i, and $\omega_{s_i}^{p_j}$ describe provider p_j's opinion towards the statement s_i.

Please note that, if we could collect external knowledge about the reliability, we could use consensus to fuse the new "external opinion" into the current $\omega_{s_i}^{model}$.

$$\omega_{s_i}^{model,\ external\ knowledge} = \omega_{s_i}^{model} \oplus \omega_{s_i}^{external\ knowledge}, \tag{5}$$

Finally, the obtained $\omega_{s_i}^{model}$ is a triple, as shown in:

$$\omega_{s_i}^{model} = \{t_{s_i}^{model}, d_{s_i}^{model}, u_{s_i}^{model}\}. \tag{6}$$

Step-4: predict statement veracity based on reliability. Step 2 and Step 3 iteratively runs until reliability scores converge. Then we could decide the statement veracity. This will be very tricky. Without historical statement labels, we could only know which statement is more reliable than another. One way to predict veracity could be select the statements whose $t_{s_i}^{model} > d_{s_i}^{model}$. Another way could be that rank all statements by their $t_{s_i}^{model}$ score, select Top N as true statements, classify the rest as false statements, and evaluate the model with eval@N, such as F1@5, Precision@10, Recall@50, and AUC@60.

However, please note that if only statement-source-value matrix is provided but without other evidences, all fancy models will degenerate into majority voting. I.E., the most popular statements are selected as true, and the most unpopular ones are taken as false; the sources belonging to majority are reliable, and the sources belonging to minority are unreliable. Therefore, we have to explore this scenario with a specific real-world problem. I will collect the problem's background knowledge, together with the statement-source-value matrix, to explore the data reliability assessment in this scenario. In this study, I will work on a real-world problem, "cancer driver gene discovery from TCGA data".

3.2 Utility case 1: cancer driver gene discovery from TCGA data

3.2.1 Background: gene mutations and cancer

Cancer is caused by gene mutations. The genes whose mutation could cause cancer are called cancer driver genes. Such genes could be classified into two broad categories:

proto-oncogenes and suppressors. Proto-oncogenes produce proteins that stimulating the cell division, and their mutation may cause ending-less cell division regardless of the need. Suppressors will produce proteins that stop cell division, and their mutation may also leads to ending-less cell division. Such ending-less cell division leads to either benign tumor, or malignant tumor (cancer). For more details, please refer [Urry et al., 2017] and CancerQuest[1].

Not all gene mutations cause cancer, and some of them may even has no effect on our body. Depending on how the gene mutates, mutations are classified into several types, such as Missense mutation, Frameshift mutation, Nonsense mutation [Cartegni et al., 2002]. In this study, I do not care specific mutation types, and depending on whether gene mutation has an effect on our body, I group mutation types to two sets: (1) first group includes “Silent mutation” and “Synonymous mutation”, which does not affect generated proteins and has no effect in our body; (2) second group includes “Missense mutation”, “Nonsense mutation” and the rest mutation types, whose mutation will change the generated proteins and then affect our body, may or may not related to cancer. For simplicity, we call first group mutation as silent mutations, and call second group as non-silent mutations.

Genes mutations that do not drive cancers as the passenger gene mutations [Wodarz et al., 2018]. Each gene has a Background Mutation Rate (BMR), which is the gene’s probability of mutating among human species, including as driver gene mutations and as passenger gene mutations [Tokheim et al., 2016]. Gene’s BMR is related to several attributes. Following [Lawrence et al., 2013], in this study, we relate gene’s BMR to (1) global expression level, (2) DNA replication time, and (3) HiC statistic, a measure of open vs. closed chromatin state (from Lieberman-Aiden et al.), and additionally, add (4) gene length.

In cell division, gene mutations randomly happens, and they will accumulate in the following cell divisions. Therefore, some cancer patients may have many mutated genes, but only a small part of them leads to cancer [Tokheim et al., 2016]. Some driver gene mutations appear in many types of cancers, and some gene mutations appear only in one type of cancer [Futreal et al., 2004, Sondka et al., 2018].

Finding the cancer driver genes is an important and difficult task in medical domain. There are an estimated 20,000-25,000 human protein-coding genes by Human Genome Project

[1] https://www.cancerquest.org/cancer-biology

in 2001 [Collins and McKusick, 2001]. Human cancer gene census in 2004 found 292 cancer driver genes [Futreal et al., 2004], and now have 724 cancer driver genes in the 2018 results [Sondka et al., 2018]. Cancer driver genes are validated in laboratories, costing a lot of resources and time. This work is still undergoing, and more cancer driver genes will be found in the future.

The Cancer Genome Atlas (TCGA) is a landmark cancer genomics program, held by National Cancer Institute and the National Human Genome Research Institute, recording the patient gene mutation data over 33 cancer types (http://cancergenome.nih.gov/). Many related studies work on cancer driver gene identification with the cancer patients' gene mutation distribution in TCGA, and hope to select the most likely cancer driver genes as inspirations for human cancer gene census institute researchers.

In this study, we will identify cancer driver genes based on the cancer patients' gene mutation distribution in TCGA. To be more specific, data sources are patients; for each gene, the corresponding statement is "gene ** is a cancer driver gene"; if a gene mutates in a patient cell samples, we say the patient provides a value "yes" to the corresponding statement. We illustrate such a problem in Figure 3. After data reliability assessed, we could rank genes by their reliability scores. I hope this study could contribute to this work.

With the above background knowledge, I conclude the following assumptions for assessing the gene's reliability being a cancer driver gene.

Assumption 3.2.1. *If a gene's mutation rate is much larger than its BMR, it is likely to be a cancer driver gene. On a specific cancer, if a gene's mutation rate is much larger than its BMR, it is likely to be a driver gene for this specific cancer.*

The most recent Human cancer gene census in 2018 [Sondka et al., 2018] found 724 cancer driver genes and there are 20,000-25,000 human protein-coding genes [Collins and McKusick, 2001]. Therefore, only a small part of human genes are related to cancer. We could further infer that, the cancer patients' gene mutation distribution is similar to the normal person's gene mutation distribution, and only a small set of cancer driver genes' mutation behavior differs. It indicates that most gene's mutation is not related to cancer, and their mutation rate in cancer patient samples should be similar to their real BMR. Therefore, it is naturally to get the above assumption about gene's mutation rate and BMR. When we calculate gene's BMR, since most genes are not cancer driver genes, we could regard the cancer driver genes as noise, and learn the BMR pattern from all genes data with Machine Learning models.

Assumption 3.2.2. *If in cancer patient samples, a gene's non-silent mutation rate is extraordinarily larger than its silent mutation rate, this gene is likely to be a cancer driver gene.*

Gene's silent mutation rate should be similar among both health people samples and among cancer patient samples, but cancer driver gene's non-silent mutation rate should be very different in both sample sets. Therefore, in cancer patient samples, the difference between regular genes' non-silent mutation rate and silent mutation rate could be learned, and if the difference is abnormally large, this gene is likely to be a cancer driver gene.

Assumption 3.2.3. *If a gene mutates frequently in cancer patient samples, this gene is very likely to be a cancer driver gene.*

It is natural to take the gene mutation frequency into consideration. If a gene never mutates in cancer patient samples, we have no evidence to suspect it as a cancer driver gene.

Assumption 3.2.4. *If a patient sample have abnormally many mutated genes (like thousands), most mutated genes should not be cancer driver genes, and we should have high uncertainty towards mutated genes in this sample data.*

As above mentioned, cancer driver genes only take a small part of the gene set. For one patient, if his/her sample has a lot of mutated genes, it is hard differentiate cancer driver genes and non-driver genes. On the contrary, if the sample only has one mutated gene, this gene is very likely to be the driver gene. We could say patient's gene mutation count increment leads to extra uncertainty.

3.2.2 Specific model design

In this subsection, we first give define several parameters, and then give the specific model design with our proposed framework.

Gene's global silent mutation rate. As above mentioned, we will use "silent mutation" to represent both silent mutations and synonymous mutations. In TCGA dataset, there are 4712 cancer patients of 21 types of cancers. Please note that, "global" means "across all types of cancer", while later "local" means "one type of cancer". Extracting silent mutations from all these samples, and have the following definition for gene g_i's global silent mutation rate:

$$s\_gmr(g_i) = \frac{\#patient\ with\ silent\ g_i\ mutations}{\#patient}, \tag{7}$$

where *#patient* is the count of patients and in this study it is 4712, and *#patient with silent* g_i *mutations* is the count of patients whose gene g_i mutates silently.

Gene's global non-silent mutation rate. Since we have already know silent mutations and synonymous mutations are not related to the cancer, we remove all these silent mutations from the dataset, and calculate each gene's global non-silent mutation rate.

$$ns\_gmr(g_i) = \frac{\#patient\ with\ nonsilent\ g_i\ mutations}{\#patient}, \tag{8}$$

where *#patient with nonsilent* g_i *mutations* is the count of patients whose $gene_j$ mutates non-silently.

Gene's BMR. Gene's *ns_gmr* is calculated based on the cancer patients datasets. If the gene is not cancer driver gene, its mutation rate should be similar to that is calculated based on normal people's data. However, the cancer driver genes' *ns_gmr* should be different from that calculated based on normal people's dataset. Therefore, we want to predict each gene's BMR. Luckily, majority genes are not cancer driver genes, and we could use the current data to train a K Nearest Neighbor (KNN) model, with gene (1) gene length, (2) global expression level, (3) DNA replication time, and (4) HiC statistic as features, and *ns_gmr* as the dependent variable. Considering the current features cannot fully describe the gene, we combine the *ns_gmr* and predicted *ns_gmr* together as the final BMR.

$$ns\_bmr(g_i) = \alpha * KNN\_ns\_gmr(g_i) + (1 - \alpha) * ns\_gmr(g_i), \tag{9}$$

where $KNN\_ns\_gmr(g_i)$ is gene non-silent mutation rated predicted by the KNN model.

When training the KNN model, we found our less than 1% features has empty value. To deal with it, for each feature, we train a KNN model with the feature as the dependent value, and *ns_gmr* as the independent value. This trained KNN model will predict the gene's feature based on gene's mutation rate.

Gene's local non-silent mutation rate. Gene's global mutation rate is calculated on the datasets containing all types of cancer. Then, for each type of cancer, we calculate the local non-silent mutation rate.

$$ns\_lmr(g_i) = \frac{\#patient\ with\ nonsilent\ g_i\ mutations}{\#patient}, \tag{10}$$

where *#patient with nonsilent g_i mutations* and *#patient* is the count of patients in a particular type of cancer.

Max gene count. As above mentioned, when a patient have too many mutated genes, figuring his/her cancer driver genes are too hard, compared with another patient who only have 2 or 3 mutated genes. Therefore, we set a bar for patient mutated gene count, *max_mutation_count*. When patient has more mutations, we have full uncertainty. We rank the patient gene mutation count from low to high, and take the mutation count at 95% as the bar. I.E., 5% patients who have more mutations will receive full uncertainty, and the 95% patients who have less mutations than *max_mutation_count* will provide information about finding cancer driver genes.

Based on above defined concepts, we design the SL based reliability representations as:

Gene background reliability. For each gene, we calculate its reliability being a cancer driver genes using the *ns_gmr*, *ns_bmr*, *ns_lmr* and *s_gmr*. We first calculate a gene's score as:

$$score'(g_i) = \frac{ns\_gmr(g_i) + e}{s\_gmr(g_i) + e} + \frac{ns\_gmr(g_i) + e}{ns\_bmr(g_i) + e} + \frac{ns\_lmr(g_i) + e}{ns\_bmr(g_i) + e}. \tag{11}$$

Please note that e is a parameter to avoid the denominator being 0, and we set it as 0.01 in this study. Score $score'(g_i)$ ranges in $(0, \infty)$, and I map it to the new range $[0, 1]$ to fit the Subjective Opinion representation with the following metrics:

$$score(g_i) = \frac{score'(g_i)}{1 + score'(g_i)} \tag{12}$$

Then we construct the gene g_i's background (BKGD) reliability as:

$$\omega_{g_i}^{BKGD} = \{t_{p_j}^{BKGD}, d_{p_j}^{BKGD}, u_{p_j}^{BKGD}\}. \tag{13}$$

where

$$\begin{cases} t_{p_j}^{BKGD} = (1-\theta)score(g_i) \\ d_{p_j}^{BKGD} = (1-\theta) * (1 - score(g_i)) \\ u_{p_j}^{BKGD} = \theta. \end{cases} \tag{14}$$

In the function, θ is a parameter to tune model's uncertainty towards such background gene reliability.

Patient's recommendation reliability. The model has a Subjective Opinion towards the patient sample. If patient have too many mutated genes, model has a high uncertainty. Model's opinion towards the patient pt_j is:

$$\omega_{pt_j}^{model} = \{t_{pt_j}^{model}, d_{pt_j}^{model}, u_{pt_j}^{model}\}, \tag{15}$$

where

$$\begin{cases} t_{pt_j}^{model} = 1 - u_{pt_j}^{model} \\ d_{pt_j}^{model} = 0 \\ u_{pt_j}^{model} = MIN(1, \frac{\#mutated\_count\_pt_j}{\#max\_mutation\_count} + e) \end{cases} \tag{16}$$

where $\#mutated\_count\_pt_j$ is the patient mutated gene count. Parameter e is added to prevent $u_{pt_j}^{model}$ being 0.

When a gene mutates in a patient's cell, we could say this patient recommend this mutated genes as the cancer driver genes. Patient pt_j's opinion towards gene g_i could directly reference model's opinion towards the gene g_i, i.e., assign $\omega_{g_i}^{model}$ to $\omega_{g_i}^{pt_j}$. Then patient's recommendation is calculated as:

$$\omega_{g_i}^{model,\ pt_j} = \omega_{pt_j}^{model} \otimes \omega_{g_i}^{pt_j}. \tag{17}$$

Gene's reliability being a cancer driver gene. Gene's final reliability being a cancer driver gene is calculated as the fusion of gene's background reliability and patients' recommendations.

$$\begin{aligned} \omega_{g_i}^{model} =& \{t_{g_i}^{model},\ d_{g_i}^{model},\ u_{g_i}^{model}\} \\ =& \omega_{g_i}^{BKGD} \oplus \omega_{g_i}^{model,\ pt_1} \oplus \omega_{g_i}^{model,\ pt_2} \oplus ... \oplus \omega_{g_i}^{model,\ pt_k}, \end{aligned} \tag{18}$$

where pt_1 to pt_k are the patients where gene g_i mutates.

3.2.3 Dataset: TCGA

The TCGA data we use is the MAF files published by TumorPortal. Totally there are data for 21 cancers, which is shown in Table 3. In the MAF files, several categories of somatic mutations are reported, including Missense and nonsense, Silent mutations, Splice site, defined as SNP within 2 bp of the splice junction, Indels that overlap the coding region or splice site of a gene or the targeted region of a genetic element of interest, Frameshift mutations, Mutations in regulatory regions.

The groundtruth is collected from TumorPortal, which gives the widely known cancer genes for 21 cancers, and the 2018 Human Cancer Gene Census. In this study, we have several evaluation metrics: (1) Mean Average Precision (MAP), (2) NDCG, (3) Area under the ranking curve (AUC), (4) Precision at different positions, and (5) Recall at different positions. Please note that, when calculate AUC, genes are sorted from most possible cancer genes at the top, and least possible cancer genes at the bottom, and sum the groundtruth cancer genes' ranking together as the AUC. Therefore, the smaller AUC, the better the model. Other evaluation metrics, higher better.

3.2.4 Baselines

Mutation Significance (MutSigCV) [Pugh et al., 2013], [Sjöblom et al., 2006] is the most frequently used tools in this area. It analyzes lists of mutations discovered in DNA sequencing, to identify genes that were mutated more often than expected by chance given background mutation processes. First, tumors are aggregated together and mutations are tallied, and then a score and p-value are calculated for each gene. A significance threshold is chosen to control the False Discovery Rate (FDR), and genes exceeding this threshold are reported as significantly mutated.

Majority is the second baseline we chose in this study, which simply counts the mutation of each gene. Though simple, for some particular cancer data, it even has better performance than MutSigCV.

In Equation 12, we proposed a score, *gene_score*, calculated with several types of mutation rates. To better evaluate our proposed Subjective Opinion based framework effective-

Table 3: Gene mutation data of 21 cancers

Cancer	Abbreviation	Patient count	Candidate genes
Acute myeloid leukemia	AML	196	2523
Bladder cancer	BLCA	99	10277
Breast cancer	BRCA	892	13430
Carcinoid	CARC	54	1232
Chronic lymphocytic leukemia	CLL	159	1831
Colorectal Carcinoma	CRC	233	14838
Diffuse large B-cell lymphoma	DLBCL	58	5190
Esophageal adenocarcinoma	ESO	141	7320
Glioblastoma multiforme	GBM	291	8195
Squamous cell carcinoma of the head and neck	HNSC	384	14427
Kidney clear cell cancer	KIRC	417	10174
Lung adenocarcinoma	LUAD	405	16862
Lung squamous cell carcinoma	LUSC	178	13692
Medulloblastoma	MED	92	824
Melanoma	MEL	118	13419
Multiple myeloma	MM	207	5602
Neuroblastoma	NB	81	1289
Ovarian cancer	OV	316	8576
Prostate cancer	PRAD	138	1745
Rhabdoid tumor	RHAB	35	181
Endometrial cancer	UCEC	248	18146

Table 4: Compare three models' performance on MAP, NDCG and AUC.

	MAP	**NDCG**	**AUC**
Majority	0.220	0.616	64391.619
MutSigCV	0.286	0.657	101612.191
MutSc	0.380	0.742	44963.286
SO-CGD	**0.397**	**0.752**	**44092.667**

ness, we propose another baseline which judge genes only with *gene_score*. If our proposed SO based model outperforms this one, it proves that Subjective Opinion based framework is effective. We call this baseline as MutSc, short for Gene Mutation Rate based Score.

Our porposed Subjective Opinion based Cancer Gene Discovery model is named as SO-CGD.

3.2.5 Experiment Results

We report the models' MAP, NDCG, and AUC results in Table 4, and report Recall and Precision at different positions in Table 5. We utilized wilxon significance test to compare the difference significance, and label the best runs with $p - value \leq 0.05$ in bold.

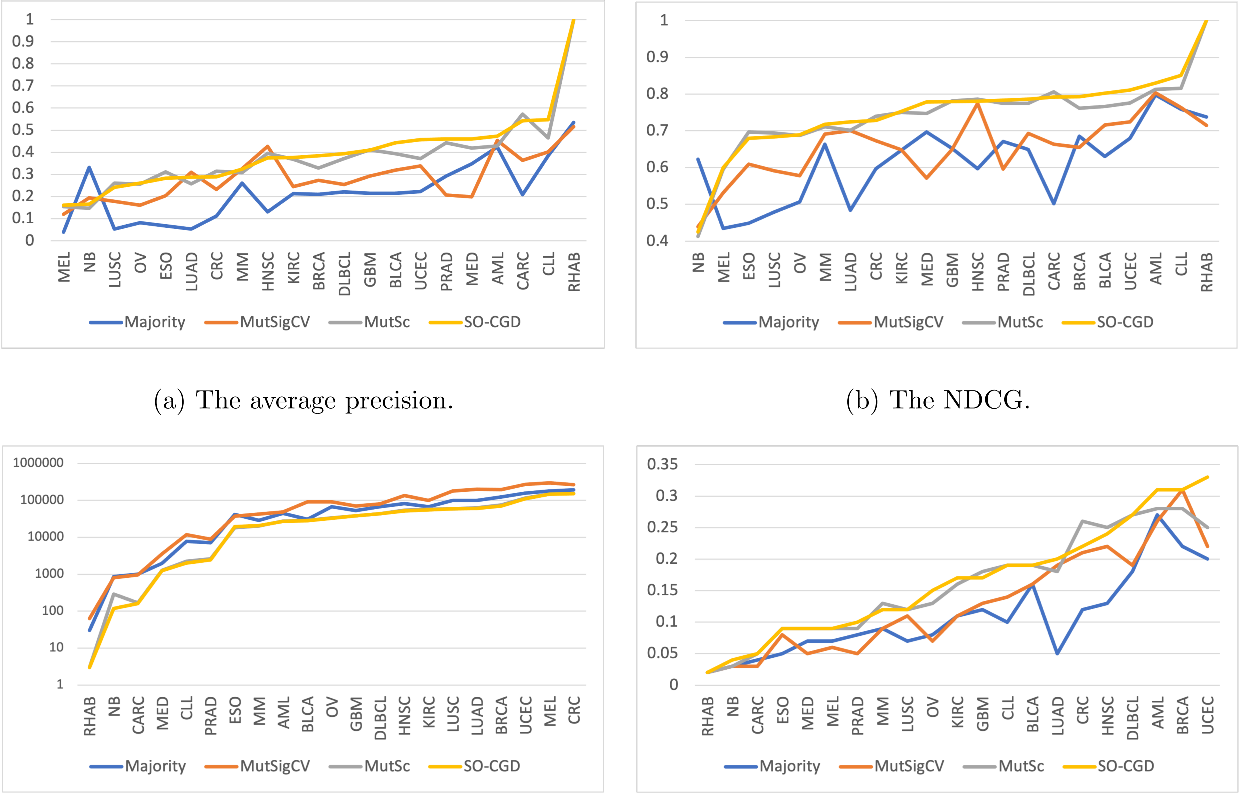
Table 4 shows that our proposed SO-CGD model outperforms three baselines by MAP, NDCG and AUC. Especially, SO-CGD and MutSc has much higher performance than Majority and MutSigCV. In terms of Recall and Precision, our proposed SOcgd and MutSc outperforms other models, and SOcgd has significant better Precision than MutSc $\{@5, @10, @20, @50, @100\}$, and significant better Recall $\{@10, @20, @100, @180\}$.

Then we compare three model's performance across different types of cancers. Figure 14(a) shows the Average Precision of three model, Figure 14(b) shows NDCG, Figure 14(c) shows AUC, and Figure 14(d) shows Precision@100. First, compare SO-CGD with Majority and MugSigCV. We could see that SO-CGD perform best on all the cancers in terms of AUC and Precision@100, while SL performs best in most cases in terms of NDCG and AP. Majority

Table 5: Precision and Recall performance at different positions of three models.

	Prec@5	**Prec@10**	**Prec@20**	**Prec@50**	**Prec@100**	**Prec@180**
Majority	0.486	0.438	0.293	0.171	0.108	0.071
MutSig	0.733	0.629	0.429	0.233	0.130	0.078
MutSc	0.810	0.652	0.488	0.267	0.159	**0.101**
SO-CGD	**0.838**	**0.695**	**0.495**	**0.271**	**0.165**	**0.100**
	Recl@5	**Recl@10**	**Recl@20**	**Recl@50**	**Recl@100**	**Recl@180**
Majority	0.119	0.196	0.231	0.341	0.407	0.463
MutSig	0.156	0.247	0.295	0.377	0.436	0.473
MutSc	**0.203**	0.285	0.367	**0.466**	0.552	0.604
SO-CGD	**0.207**	**0.297**	**0.373**	**0.462**	**0.578**	**0.612**

voting, though the strategy is naive, is quite effective in the NB cancer. Also, MutSigCV has better AP and NDCG on LUAD and HNSC. But our proposed SO-CGD model significantly outperforms both baselines. Then, we compare SO-CGD and MutSc. In most cancers data, SO-CGD provides a similar but slightly better performance than MutSc. It indicates that, our proposed *gene_score* computed with the gene's various mutation rate is very effective. Then SO-CGD's slight but significant improvement over MutSc indicates the effectiveness of our proposed Subjective Opinion based data reliability assessment framework.



(a) The average precision.

(b) The NDCG.

(c) The AUC.

(d) The Precision@100.

Figure 14: Compare three models' performance on each type of cancer.

3.3 Summary

In this chapter, we have presented the Subjective Opinion based data reliability assessment framework in the scenario where there is neither labeled training data nor conflicting values. This is the most challenging research question in this dissertation. We first propose a general framework, and then give specific designs with a real-world problem, finding cancer driver genes with TCGA data. Experiment results shows that our proposed models significantly outperforms state-of-art baseline models, validating the effectiveness of our proposed framework.

4.0 RQ 2: How to assess data reliability when sources provide consistent values and historical statement labels are provided

Compare with Research Question 1, this chapter explores a relatively easier scenario, where a set of labeled training data is provided to help data reliability assessment. Providers still provide non-conflicting values to the statements, and hence when a source provide value to a statement, we could say that the source votes the statement. In the first section, we will introduce the proposed Subjective Opinion based framework for data reliability assessment, and then in the second section, the proposed framework is used in a real-world problem.

4.1 Subjective Opinion based data reliability assessment

Subjective Opinion based data reliability assessment framework for this scenario is basically same as that in Scenario 1. The only difference is that labeled training data (i.e., statement veracity) is provided in this scenario. Intuitively, we should have following assumptions:

Assumption 4.1.1. *If a provider votes many reliable statements, then this provider is also reliable; if provider votes many unreliable statements, then this provider should also be unreliable.*

Assumption 4.1.2. *If a provider votes many statements, but statements veracity can not be declared, then we should have high uncertainty towards the provider.*

Assumption 4.1.3. *If a statement receives many votes from highly uncertain providers, then we should keep high uncertainty towards this statement.*

Related works talked in Section 2.3 only consider the first assumption, and usually ignore the uncertainty in the latter two assumptions. Our framework take all of them into consideration.

The framework first prepossesses the dataset, initialize the statement veracity labels,

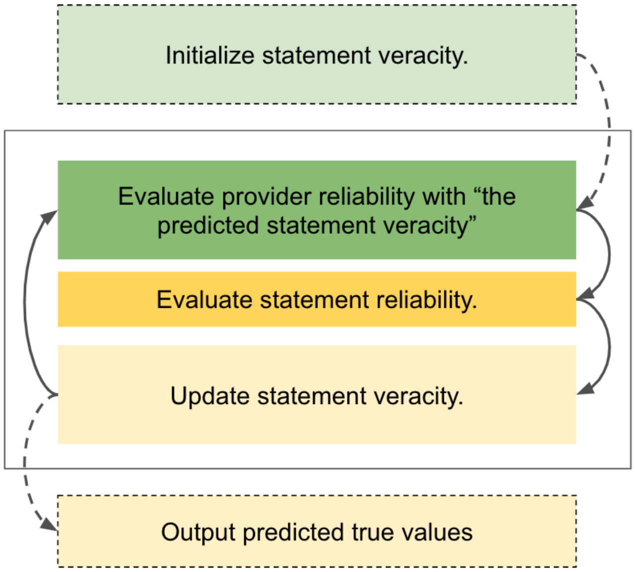

Initialize statement veracity.
Evaluate provider reliability with "the predicted statement veracity"
Evaluate statement reliability.
Update statement veracity.
Output predicted true values

Figure 15: Subjective Opinion based data reliability assessment framework for Scenario 3.

then iteratively assess provider/source reliability ($\omega_{p_j}^{model}$), statement reliability ($\omega_{s_i}^{model}$) and predict statement veracity based on reliability until converge, as shown in Figure 15. In this scenario, labeled training data (i.e., statement veracity) is provided, and therefore, we could directly use the provided statement veracity labels. Also, with labeled training data, assessed reliability and predicted statement veracity should be more accurate than that in Scenario 1. Therefore, we include "statement veracity prediction" as a part of iteration, and predicted/provided statement veracity is used in provider reliability assessment.

Step-1: preprocessing the dataset and initialize statement veracity. In the preprocessing procedure, if a provider p_j only votes one statement s_i, this provider p_j should be removed. The reason is that (1) if the statement s_i veracity is unknown, we cannot infer p_j's reliability, and hence p_j cannot provide information; (2)if the s_i's veracity is known, as p_j does not vote other statements, model needs no information from it.

For the labeled statements, their ground-truth veracity is directly assigned, and will not be changed in the following steps. For the unlabeled statements, their veracity is labeled as "unknown", and will be continuously updated in each iteration with following steps.

Step-2: evaluate provider reliability. Provider reliability is defined as model's opinion towards the statement "provider p_j is reliable", which is defined as:

$$\omega_{p_j}^{model} = \{t_{p_j}^{model}, d_{p_j}^{model}, u_{p_j}^{model}\}. \tag{19}$$

Since historical statement labels are provided, then we could define

$$\begin{cases} t_{p_j}^{model} = \frac{rc(p_j)}{rc(p_j)+fc(p_j)+uc(p_j)} * (1-\theta), \\ d_{p_j}^{model} = \frac{fc(p_j)}{rc(p_j)+fc(p_j)+uc(p_j)} * (1-\theta), \\ u_{p_j}^{model} = \frac{uc(p_j)}{rc(p_j)+fc(p_j)+uc(p_j)} * (1-\theta) + \theta, \end{cases} \tag{20}$$

where $uc(p_j)$ is the count of statements without labels, to which this source p_j provides values; $rc(p_j)$ is the count of true statements (i.e., ground-truth true or predicted true), to which this source p_j provides values; $fc(p_j)$ is the count of false statements (i.e., ground-truth false or predicted false), to which this source p_j provides values. Please note that, in the first iteration, $rc(p_j)$ and $fc(p_j)$ comes from labeled training data, and in later iterations, they also include the predicted statement veracity. Parameter θ represent model's basic uncertainty towards the providers, and could be tuned with the dataset.

Step-3: evaluate statement reliability. For each statement, we collect all sources who provide values to it. Based on whether provider has same belief in all statements that he provides values to, we have two different statement reliability assessment cases as RQ1.

Case 1 If provider/source p_j has same belief in all statements that he provide values to, then directly fuse these sources' reliability with Subjective Logic consensus operation. The fused opinion represents our belief for the statement "statement s_i is real".

$$\omega_{s_i}^{model} = \omega_{p_1}^{model} \oplus \omega_{p_2}^{model} \oplus ... \oplus \omega_{p_k}^{model}, \tag{21}$$

where p_1, p_2,..., p_k are sources providing values to s_i.

Case 2 If provider/source p_j has different beliefs in the statements that he provide values to, first get the Subjective Opinion to statements according to provider's recommendation, and then fuse these opinions with Subjective Logic consensus operation. The fused opinion represents our belief for the statement "statement s_i is real".

$$\omega_{s_i}^{model} = (\omega_{p_1}^{model} \otimes \omega_{s_i}^{p_1}) \oplus (\omega_{p_2}^{model} \otimes \omega_{s_i}^{p_2}) \oplus ... \oplus (\omega_{p_k}^{model} \otimes \omega_{s_i}^{p_k}), \tag{22}$$

where p_1, p_2,..., p_k are sources providing values to s_i, and $\omega_{s_i}^{p_j}$ describe provider p_j's opinion towards the statement s_i.

Finally, the obtained $\omega_{s_i}^{model}$ is a triple, as shown in:

$$\omega_{s_i}^{model} = \{t_{s_i}^{model}, d_{s_i}^{model}, u_{s_i}^{model}\}. \tag{23}$$

Step-4: predict statement veracity based on reliability. Since we have historical statement labeling data, and the $\omega_{s_i}^{model}$ has been predicted, we could train a classification model, such as SVM, to automatically learn and predict statement veracity. I.e., $\{t_{s_i}^{model}, d_{s_i}^{model}, u_{s_i}^{model}\}$ are features, and statement veracity labels are target classes. These three steps iteratively run until converge.

4.2 Utility case 2: fake news detection

4.2.1 Dataset: FakeNewsNet

FakeNewsNet [Shu et al., 2018] is selected as the validation dataset in this study. It consists of two real-world datasets, BuzzFeed and PolitiFact. BuzzFeed contains 90 real news and 90 fake news, with 15,257 users interacting with (re-tweet or like) the news. PolitiFact contains 120 real news, 120 fake news, and 23,865 users interacting with the news. Following the first step in the framework, we remove the users who shares only one news article. After preprocessing, 3,002 users are left in BuzzFeed, and 4,139 users are left in PolitiFact. Also, we find the PolitiFact data is denser than BuzzFeed data, i.e., people share more news in PolitiFact.

Please note that if a news article is not shared by anyone, or shared by only one user but the user's reliability could not be assessed (removed in preprocessing), we directly label the news as fake news and do not update its label in iterations. We do know such a straightforward strategy leads to failure, but such loss on unpopular news is acceptable in real-world scenarios.

4.2.2 Experiment settings

Following procedures in [Shu et al., 2017b], [Tacchini et al., 2017], and [Della Vedova et al., 2018], we learn and evaluate our models with 5-fold cross-validation, i.e., 20% of data is used as testing, while 80% of data is used to train the model. Each cross-validation is repeated 50 times, and the average performance with standard deviation is reported. Accuracy, Precision, Recall, and F1 of detecting fake news are selected as the evaluation metrics.

Prob_fnd has no parameters, while SO_fnd has one parameter α, which describes people's natural/basic uncertainty. Following [Shu et al., 2017b], [Tacchini et al., 2017] and [Della Vedova et al., 2018], we select $\alpha = 0.9$, because it achieves the highest performance with both datasets in cross-validation.

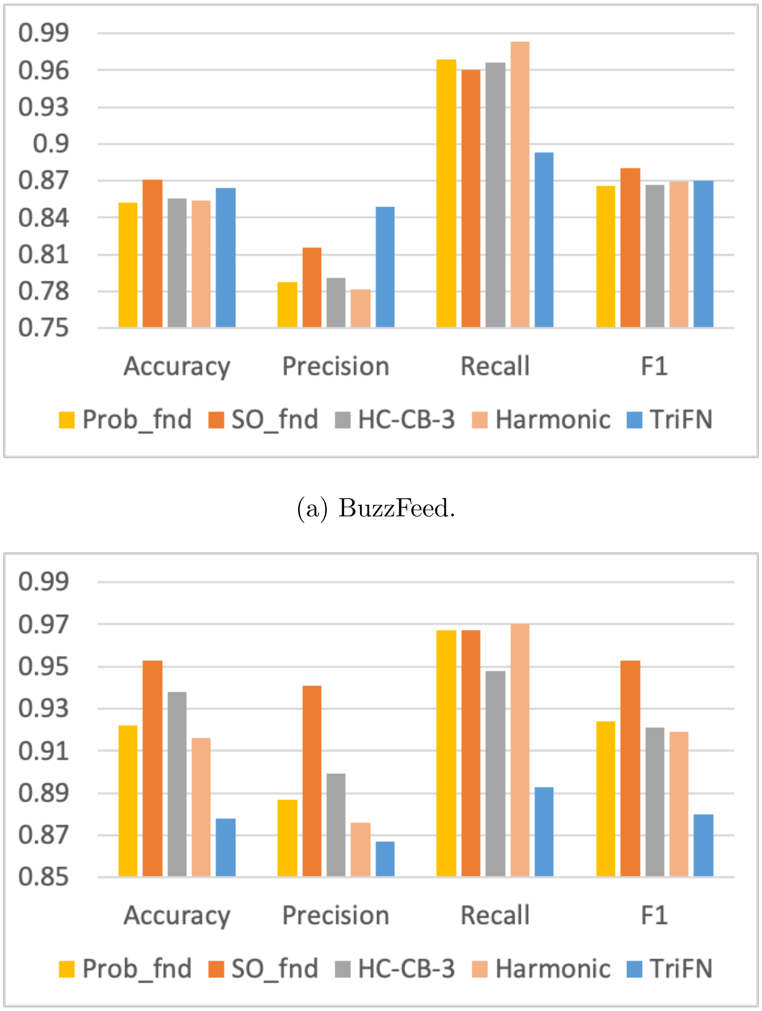
4.2.3 Baselines

Harmonic from [Tacchini et al., 2017]. This method is very similar to our proposed Prob_fnd, iteratively evaluating the reliability scores of both users and news, and both methods ignore the unknown news and users in the calculation. The major difference is that Harmonic explicitly differentiate reliable users from unreliable users, and real news is those that accumulate more scores from reliable users, while fake news is those that accumulate more scores from unreliable users. On the other hand, in Prob_fnd, news reliability is defined as the average reliability of the users that shared it without explicitly different reliable and unreliable users.

HC-CB-3 from [Della Vedova et al., 2018]. This method is developed based on Harmonic. It utilizes the word-level features of news content with a logistic regression model. If the news is shared by more than λ people, social-network based Harmonic is used; otherwise, content based classification is used.

TriFN from [Shu et al., 2017b]. This method designed a Tri-Relationship embedding framework, which utilizes the information from news content, news-user interaction, and news-publisher relationship. TriFN shows much better performance than several other baselines, which use content based or social-network based features, and they are not included in this paper due to page limit.

Prob_fnd from [Zhang and Zadorozhny, 2020]. This method martians reliability score for both provider and statements, and runs iteratively. The main difference between Prob_fnd and SO_fnd is that Prob_fnd doesn't explicitly record the uncertainty in its reliability scores. Comparing Prob_fnd and our SO_fnd, we could clearly see the effectiveness of Subjective Opinion based representations.



(a) BuzzFeed.

(b) PolitiFact.

Figure 16: Repeated 5-fold cross-validation results on two real-world datasets.

Table 6: Repeated 5-fold cross-validation results on two real-world datasets.

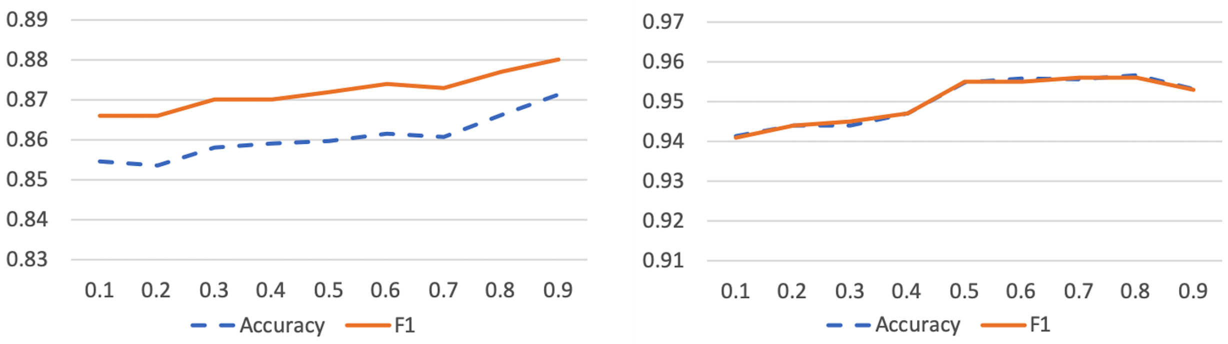
		Prob_fnd	SO_fnd	HC-CB-3	Harmonic	TriFN
Buzz- Feed	ACC	.852 ± .055	.871 ± .051**	.856 ± .052	.854 ± .052	.864 ± .026*
	Prec	.788 ± .086	.816 ± .079*	.791 ± .076	.782 ± .075	.849 ± .040**
	Recl	.969 ± .043*	.960 ± .004	.966 ± .045	.983 ± .041**	.893 ± .013
	F1	.866 ± .052	.880 ± .050**	.867 ± .050	.869 ± .050	.870 ± .019*
Polit- iFact	ACC	.922 ± .036	.953 ± .029**	.938 ± .029*	.916 ± .042	.878 ± .020
	Prec	.887 ± .056	.941 ± .048**	.899 ± .057*	.876 ± .074	.867 ± .034
	Recl	.967 ± .034*	.967 ± .034*	.948 ± .046	.970 ± .030**	.893 ± .023
	F1	.924 ± .035*	.953 ± .030**	.921 ± .041	.919 ± .044	.880 ± .017

x^{**}: the run with the best performance.
x^{*} : the run with the second best performance.

4.2.4 Experiment Results

Experiment results are shown in Table 6, and their comparison is better illustrated in Figure 16. Best and second-best performed runs are labeled with '**' and '*'. Our proposed SO_fnd has the best performance on Accuracy and F1 in both datasets, indicating that SO_fnd can differentiate fake news from real news much more accurately than other methods. Further, it beats Prob_fnd in almost every evaluation metrics, showing that keeping a record of unknown cases as an uncertainty value is essential, and can highly improve the fake news detection accuracy. On the other hand, we can find that Prob_fnd has very similar performance with Harmonic, implying that whether or not explicitly differentiate reliable users from unreliable users does not make a big difference in these two datasets.

Also, both baselines and our proposed methods have lower precision and a higher recall on both datasets. It indicates that most fake news is detected, but much real news is wrongly classified as fake news. Such a model is better than one that wrongly classifies fake news to real news. Because (1) the broad propagation of fake news may lead to inestimable damages,



(a) BuzzFeed.

(b) PolitiFact.

Figure 17: Accuracy and F1 of SO_fnd varying with different α on BuzzFeed and PolitiFact.

but people can search online if a piece of real news is filtered but needed; (2) we can hire people to manually check and filter real news from these automatically predicted fake news, and as the data size decrease a lot, the manual effort cost is smaller.

To be more specific, on the BuzzFeed dataset, SO_fnd got the best Accuracy 87.1% and F1 score 88.0%, while TriFN got second-best accuracy and F1. In terms of precision and recall, we can see that SO_fnd has second-best precision 81.6%, and Prob_fnd has second-best recall 96.9%. On the PolitiFact dataset, SO_fnd has the best Accuracy 95.3%, F1 score 95.3%, precision 94.1%, and second-best recall 96.7%; while Prob_fnd gained second best F1 score 92.4%, and recall 96.7%. Also, TriFN and our proposed SO_fnd have a relatively more balanced precision and recall than other baselines, but SO_fnd have better performance than TriFN across almost all evaluation metrics. HC-CB-3, Harmonic, and Prob_fnd are more imbalanced, sacrificing the precision to get the higher recall, but the F1 score is still less than SO_fnd.

4.2.5 Further Discussion

Is SO_fnd performance sensitive to natural uncertainty parameter?

SO_fnd has one parameter α, which describes people's natural/basic uncertainty. Above reported results are obtained when $\alpha = 0.9$, which is the highest performance with both datasets in cross-validation. We report the Accuracy and F1 of SO_fnd when α changes in range $[0.1, 0.9]$. We repeat the cross-validation procedure 50 times, and the average performance is reported in Figure 17.

From Figure17, we can observe that though SO_fnd prefers larger α in both datasets when α varies across $[0.1, 0.9]$, the accuracy and F1 do not change a lot, with the increments less than 2%. It shows SO_fnd is relatively stable to the parameter α in range $[0.1, 0.9]$.

How does the training data size affect the performance of the methods?

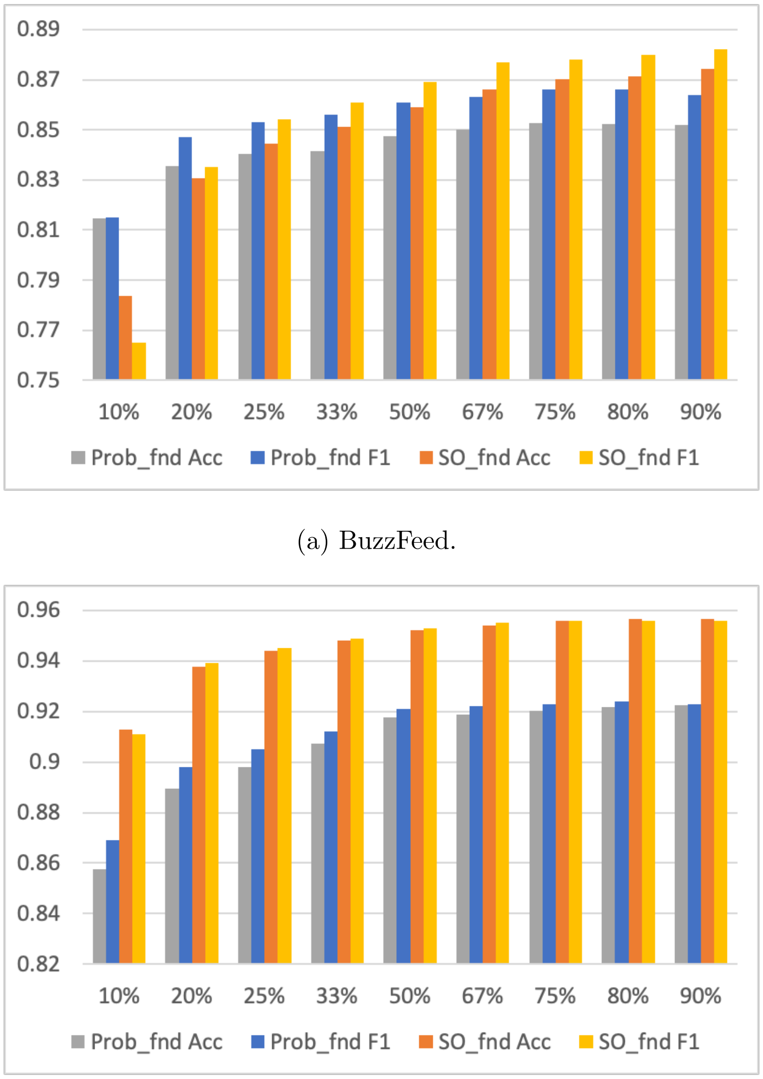
In this subsection, we explored the performance of Prob_fnd and SO_fnd when they are trained by different sizes of data. As shown in Figure 18, the size of training data increases from 10% to 90%, and the accuracy is evaluated for each model on both BuzzFeed and PolitiFact. We repeat the training and testing procedure 50 times for each run, and the average performance is reported.

From Figure 18, we can observe that, Accuracy and F1 score of Prob_fnd and SO_fnd all increase in both datasets when training data size rises. Also, SO_fnd's performance outperforms Prob_fnd in most cases, except when they are trained with 20% or less data on BuzzFeed.

Shu et al. reported TriFN's performance with the different training set sizes in [Shu et al., 2017b]. On BuzzFeed, when training data is 40% and less, TriFN's Accuracy and F1 is less than 80%; however, Prob_fnd's Accuracy and F1 are above 80% even with 10% training data, and SO_fnd's Accuracy and F1 are above 80% with 20% or more training data. On PolitiFact, TriFN's Accuracy and F1 are above 80% with 40% or more training data; however, SO_fnd's Accuracy and F1 are above 90% even with 10% training data, and Prob_fnd's Accuracy and F1 are above 90% with 33% or more training data. It shows that, compared to TriFN, our proposed two models are able to achieve a similar or even better performance with much less labeled training data.

Does users voting less news articles should receive higher uncertainty?

Intuitively, if we observe user A forwarding 100 news articles, and user B only forwarding 3 news articles, we should be more confident about our judgement towards user A than that towards user B. I.E., if provider p_j gives values to a few statements, then we could only



(a) BuzzFeed.

(b) PolitiFact.

Figure 18: Accuracy and F1 of Prob_fnd and SO_fnd varying with different training data size on BuzzFeed and PolitiFact.

Table 7: Compare the SO_fnd and the extended version on BuzzFeed and PolitiFact.

BuzzFeed	SO_fnd	Extended version	PolitiFact	SO_fnd	Extended version
accuracy	0.871	0.874	accuracy	0.953	0.955
precision	0.816	0.817	precision	0.941	0.946
recall	0.96	0.966	recall	0.967	0.967
f1	0.88	0.879	f1	0.953	0.955

collect a small set of evidence, and hence model's uncertainty should be higher. Therefore, I enrich the Equation 20 with function $f(p_j)$ to consider the size of statements that provider p_j votes.

$$\begin{cases} t_{p_j}^{model} &= \frac{rc(p_j)}{rc(p_j)+fc(p_j)+uc(p_j)} * (1 - f(p_j)), \\ t_{p_j}^{model} &= \frac{fc(p_j)}{rc(p_j)+fc(p_j)+uc(p_j)} * (1 - f(p_j)), \\ t_{p_j}^{model} &= \frac{uc(p_j)}{rc(p_j)+fc(p_j)+uc(p_j)} * (1 - f(p_j)) + f(p_j), \end{cases} \tag{24}$$

where function $f(p_j)$ is defined as:

$$f(p_j) = \theta * (1 - \frac{rc(p_j) + fc(p_j) + uc(p_j)}{1 + max(\{\#rc(p_k) + fc(p_k) + uc(p_k)|p_k \in \{P\}\})}). \tag{25}$$

Parameter θ is used to tune the weight of this size effect.

Table 7 shows the extended model's performance. Though most evaluation metrics increase with extended model, there is no significant difference ($p - value > 0.05$). Therefore, at least on these two real-world dataset, when calculate provider's reliability, taking the size of statements that provider votes into consideration is not essential.

In what situations, will Prob_fnd and SO_fnd win and fail?

In this subsection, we explore in what situations two proposed methods shall win and shall fail. This experiment is conducted on PolitiFact because two methods' performance difference is more significant on it than BuzzFeed.

First, we use all news labels in PolitiFact to calculate *user_reliability* in Formula 1. Second, we mark smart users are those whose *user_reliability* $>= 0.8$, mark credulous users

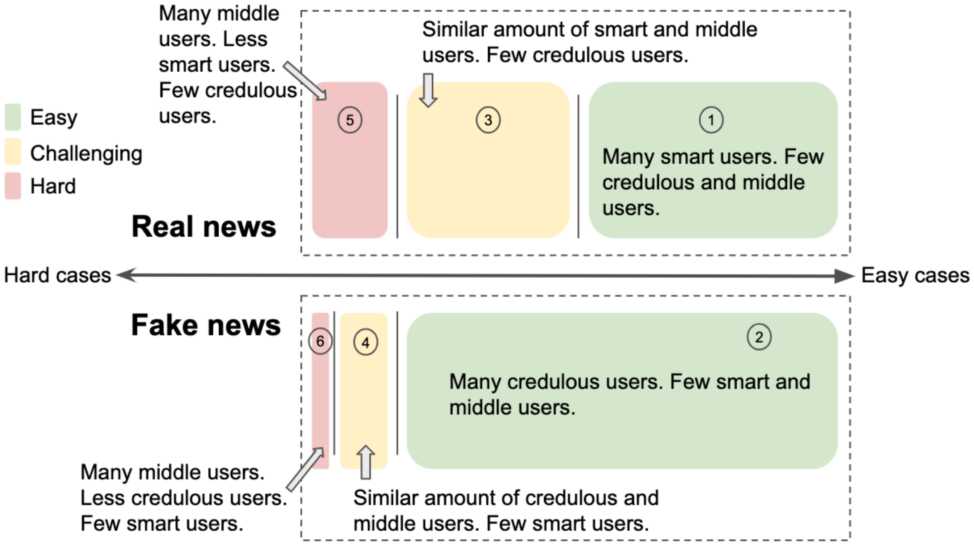

Easy
Challenging
Hard
Many middle users. Less smart users. Few credulous users.
Similar amount of smart and middle users. Few credulous users.
5
3
1
Many smart users. Few credulous and middle users.
Real news
Hard cases
Easy cases
Fake news
6
4
2
Many credulous users. Few smart and middle users.
Many middle users. Less credulous users. Few smart users.
Similar amount of credulous and middle users. Few smart users.

Figure 19: Easy classified, challenging, and hard classified news in PolitiFact.

are those whose *user_reliability* <= 0.2, and mark other users as the middle users who are neither so smart nor so credulous. Third, as shown in Figure 19, based on the distribution of three types of users, we identify three levels of difficulty for news classification:

- **Easily classified news.** News that is shared mainly by smart users, as shown in area 1 in Figure 19, are very likely to be real news; news that is shared mainly by credulous users, as shown in area 2, are very likely to be fake news. Smart and credulous users could be easily identified with training data, and hence models can easily classify them.
- **Challenging news.** If the news is shared by a similar amount of middle users and smart/credulous users, as shown in areas 3 and 4 in Figure 19, the classification performance is affected by the reliability assessment of middle users. The reliability assessment accuracy for middle users vary in different models, and the classification of such news is challenging. We found that our proposed two models were able to identify them successfully.
- **Hard classified news.** If the news is shared mainly by middle users, as shown in area 5 and 6 in Figure 19, it is hard to classify them. The classification performance is directly

decided by (1) the reliability assessment accuracy for the middle users, and (2) how the news reliability assessment is designed. We checked the failure cases for both methods in repetitive experiments and found that there were 15 out of 17 frequently appearing failure cases in Prob_fnd, and 7 out of 9 frequent failure cases in SO_fnd, can be attributed to a large number of middle users, and SO_fnd has relatively fewer losses than Prob_fnd.

Two other failure cases for Prob_fnd and SO_fnd are: (1) a piece of real news (news id 84) is spread by more credulous users and hence is wrongly classified. (2) unknown news, whose related user information (reliability) cannot be assessed from the data, is directly classified as fake news. Hence, a piece of real news (news 22) is wrongly classified.

As shown in Figure 19, area 3 and 5 are much larger than area 4 and 6, as we found that in PolitiFact, most fake news is easily classified news and only a few are challenging or hard classified fake news, while nearly half of real news is challenging or hard to be classified. It explains that, when used as the SVM classification feature in first iteration of the SO_fnd, *news_distrust* has a better performance than *news_trust*, and hence is selected. When *news_trust* is high, the news are likely to be real, and when *news_distrust* is high, the news are likely to be fake. However, when *news_trust* is low, the news may be hard classified real news, or fake news; when *news_distrust* is low, the news is very likely to be real (because challenging and hard classified fake news are too few). Hence, we use *news_distrust* as the SVM feature for classification.

4.3 Summary

In this chapter, we have presented the Subjective Opinion based data reliability assessment framework in the scenario where labeled training data is provided, which provide rich evidences for model to accurately make assessment. We validate our framework on a real-world dataset. Experiment results shows that our proposed models significantly outperforms state-of-art baseline models, validating the effectiveness of our proposed framework.

5.0 RQ 3: How to assess data reliability when source provide conflicting values without historical data

In this chapter, we will discuss the data reliability assessment in the scenario where historical data are not provided, but the conflicting values may appear. This question is supposed to be moderate challenging, as conflicting values could provide additional evidences for reliability assessment.

5.1 Subjective Opinion based data reliability assessment

In this Subjective Opinion based data reliability assessment framework, provider reliability is represented by $\omega_{p_j}^{model}$, describing the model's opinion towards the declaration "provider p_j is reliable". Different from Scenario 1 and 2, in this scenario, one statement may receive several different values from multiple providers, and hence we evaluate the reliability of candidate value v instead of statement s_i, which is represented by ω_v^{model}. In such a scenario, I have following assumptions and the proposed framework consider them all:

Assumption 5.1.1. *If values from a provider are usually (or quite close to) true values, this provider is reliable; if provider's values are far from the truth, then this provider is very unreliable.*

Assumption 5.1.2. *Assume most data are real, then if a statement's candidate value distribution is particularly scattered (or do not support each other's existence), then it is harder to predict the truth than anther statement with concentrated distribution. Evidences from the former statement should receive higher uncertainty than the latter one.*

Assumption 5.1.3. *If a provider provides values to many statements, with the evidences accumulating, we should have a lower uncertainty towards the provider.*

Assumption 5.1.4. *Also, we should consider the basic uncertainty that even the most reliable provider/statement/value could be wrong.*

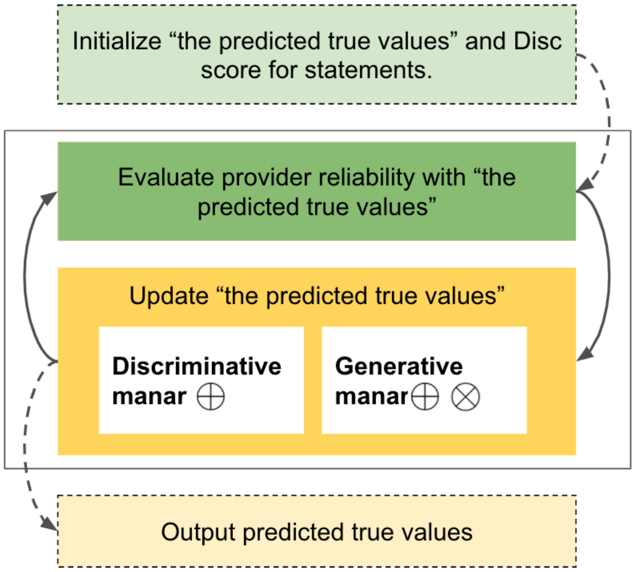

Initialize “the predicted true values” and Disc score for statements.
Evaluate provider reliability with “the predicted true values”
Update “the predicted true values”
Discriminative manar ⊕
Generative manar⊕ ⊗
Output predicted true values

Figure 20: Subjective Opinion based data reliability assessment framework for Scenario 3.

My framework first initialize the statement true values and discrimination scores, then iteratively assess provider/source reliability and update the statement true value until converge, as shown in Figure 20.

Step-1: initialize the statement true values and statement discrimination score. Before entering iteration, all statements are initialized with a true value $v_{i'}$, which is obtained with naive strategies, such as Majority Voting, Average, Maximum, Minimum, or other algorithms. Later, $v_{i'}$ will be updated in each iteration.

Then, each statement is assigned a discrimination score. This score ignores the provider's reliability (i.e., equally treat every provider), only focus on statement's candidate value distribution. Given an statement s_a, if most of the candidate values are very similar/close to each other, the true value is very likely to be one of them or very close to them. However, given another statement s_b, if candidate values vary largely, even for a human, it is hard to infer the true value, and the inferred score for each value is less unconvincing. When evaluate a provider's reliability, the statement s_a can provide a more convincing evidence than the statement s_b. Based on such consideration, we propose use **statement's discrimination**

score to describe “referencing this statement’s data, model’s ability to differentiate the reliable and unreliable providers”. Statement s_i’s discrimination score $Disc(s_i)$ is defined as:

$$Disc(s_i) = \frac{\sum_{x=1}^{m}\sum_{y=1}^{m}\{Imp(v_{ix} \to v_{iy})|x \neq y\}}{\sum_{x=1}^{m}\sum_{y=1}^{m}\{1|x \neq y\}} \tag{26}$$

where $Imp(v_{ix} \to v_{iy})$ reflects the implication from v_{ix} to v_{iy}, introduced from [Yin et al., 2008]. It is a value reflecting to what degree V_{iy} is (partially) true if V_{ix} is correct. In this study, $Imp(v_{ix} \to v_{iy})$ ranges from 0 to 1, with 0 means no such implication, and 1 means v_{ix} fully support V_{iy} is true. Its formula should be defined in specific tasks. We need consider the task background, and also consider value’s meaning, and value’s types.

Step-2: calculate provider’s reliability. Provider’s reliability is defined as our opinions towards the statement “provider p_j is reliable”, which is define as:

$$\omega_{p_j}^{model} = \{t_{p_j}^{model}, d_{p_j}^{model}, u_{p_j}^{model}\}. \tag{27}$$

Intuitively, if the provider’s values are close to the statement true values, then this provider is reliable; otherwise, this provider is unreliable. Therefore, we define:

$$\begin{cases} t_{p_j}^{model} = \frac{\sum_{i=1}^{n}\{Disc(s_i)*Imp(v_{i'} \to v_{ij})|v_{ij} \neq null\}}{\sum_{i=1}^{n}\{Disc(s_i)|v_{ij} \neq null\}} * (1 - u_{p_j}^{model}) \\ d_{p_j}^{model} = 1 - t_{p_j}^{model} - u_{p_j}^{model} \\ u_{p_j}^{model} = \theta * \frac{\sum_{i=1}^{n}\{1 - Disc(s_i)|v_{ij} \neq null\}}{\sum_{i=1}^{n}\{1|v_{ij} \neq null\}} + \alpha, \end{cases} \tag{28}$$

where α is a parameter representing our basic uncertainty, θ is a parameter tunning the weight of averaged $Disc$ score, and $Imp(v_{i'} \to v_{ij})$ is the implication from the predicted true value to v_{ij}, which is same as the one in Equation 26, defined based on specific task background.

Step-3: update the true values. Each statement’s true value could be decided based on each value’s reliability, either in a discriminative way, or in a generative way. If values are numeric, we could select the value with highest reliability as the true value, and could also calculate a true value based on the reliability score. If values are categorical, then we will have to chose one from the existing candidate values in a discriminative way.

Case 1: in a discriminative manner. If provider/source p_j has same belief in all values that he provides, then provider p_j's reliability could represent provider p_j's recommendation towards the value. One value could be provided by several providers, and then fuse these sources' reliability with Subjective Logic consensus operation. The fused opinion represents our belief for the statement "v_{ij} is the true value for statement s_i", which is represented as:

$$\omega_{v_{ij}}^{model} = \{t_{v_{ij}}^{model}, d_{v_{ij}}^{model}, u_{v_{ij}}^{model}\}. \tag{29}$$

It is obtained by:

$$\omega_{v_{ij}}^{model} = \omega_{p_1}^{model} \oplus \omega_{p_2}^{model} \oplus ... \oplus \omega_{p_k}^{model}, \tag{30}$$

where p_1, p_2,..., p_k are providers who provide values for s_i. Given a statement s_i, for each candidate value $\{v_{ij}|j = 1, ..., m,\ v_{ij} \neq null\}$, we compare their reliability $\omega_{v_{ij}}^{model}$, and select the value with highest trust $t_{v_{ij}}^{model}$ as the true value.

Case 2: in a generative manner. In this case, for the statement s_i, we assume that value $v_{i''}$ is the temporary true value. Then provider p_j gives an opinion towards "the value $v_{i''}$ is the true value for statement s_i" based on the distance between his provided value v_{ij} and this $v_{i''}$. The model could learn $v_{i''}$'s reliability based on p_j's recommendation. After considering statement s_i all providers' recommendation, the model could get a fused idea about this temporary true value $v_{i''}$, and based on which generates a predicted true value.

There are many ways to obtain the temporary true value $v_{i''}$, and we recommend to use statement's maximum or minimum candidate value. Take maximum for example, after obtaining the final opinion towards this $v_{i''}$, if trust is high, the generated value should approach the maximum candidate value; if trust is low, the generated value should approach the minimum candidate value. Vice versa. On the other hand, if chose median, average, or majority, when the trust in the final fused opinion is low, the model has no idea about whether the predicted true value should be bigger or lower than $v_{i''}$. In this study, for each statement s_i, we chose its maximum candidate value as the temporary true value, $v_{i''} = max(\{v_{ij}|j \in \{1, \ldots, m\})$.

First, on each statement s_i, we normalize all the candidate values in the following manner:

$$v_{ij}^{'} = \frac{v_{ij} - min(\{v_{ij}|j \in \{1, \ldots, m\})}{max(\{v_{ij}|j \in \{1, \ldots, m\}) - min(\{v_{ij}|j \in \{1, \ldots, m\})}, \tag{31}$$

Then we have $v_{ij}^{'} \in [0, 1]$. After this, statement "true value of statement s_i is the max candidate value $max(\{v_{ij}|j \in \{1, \ldots, m\})$" is mapped to "in the normalized space, true value of statement s_i is 1". Thereby, the provider p_j's opinion towards the statement can be defined as:

$$\omega_{v_{i''}=1}^{p_j} = \{(1 - \beta)v_{ij}^{'}, 1 - (1 - \beta)v_{ij}^{'} - \beta, \beta\}, \tag{32}$$

where β describe provider's fundamental uncertainty, similar to above α.

Second, the provider p_j can recommend his opinion $\omega_{v_{i''}=1}^{p_j}$ to the model. Recommendation operation can help people know the statement according to their acquaintances. Thus, model's opinion towards "in the normalized space, true value of statement s_i is 1" could be obtained:

$$\omega_{v_{i''}=1}^{model,p_j} = \omega_{p_j}^{model} \otimes \omega_{v_{i''}=1}^{p_j}. \tag{33}$$

Statement s_i has a set of candidate values from several providers $\{p_1, p_2, ..., p_k\}$, and the model should have a summarized opinion based on all recommendations. Consensus operation can help fuse several opinions towards one statement together. The model's final opinion towards the temporary true value $v_{i''}$ is defined as:

$$\omega_{v_{i''}=1}^{model,p_1,p_2,...,p_k} = \omega_{v_{i''}=1}^{model,p_1} \oplus \omega_{v_{i''}=1}^{model,p_2} \oplus ... \oplus \omega_{v_{i''}=1}^{model,p_k}. \tag{34}$$

In the final fused opinion $\omega_{v_{i''}=1}^{model,p_1,p_2,...,p_k}$, the trust reflects model's confidence about the temporary true value $v_{i''} = 1$ in the normalized space. Map $t_{v_{i''}=1}^{model,p_1,p_2,...,p_k}$ to the original numerical space, and we could get the predicted true value $v_{i'}$.

$$\begin{aligned} v_{i'} = &t_{v_{i''}=1}^{model,p_1,p_2,...,p_k} * (max(\{v_{ij}|j \in \{1, \ldots, m\}) - min(\{v_{ij}|j \in \{1, \ldots, m\}))+ \\ &min(\{v_{ij}|j \in \{1, \ldots, m\}). \end{aligned} \tag{35}$$

5.2 Utility case 3: Find true book author list

5.2.1 Dataset: Book

It is a popular categorical dataset in truth discovery area, composed by Luna Xin Dong[1]. Its data describes that for each book, online bookstores post author list in their web pages, but some data is wrong. It contains the information on ISBN, book name, authors, online bookstore name for 1265 books. Totally, there are 894 bookstores and they generate 26,494 author lists.

We have two testing data. First one is the gold testing dataset published by Luna Dong, consisting of 100 books. The second testing dataset is composed of 161 book, containing the first 100 books and other 61 books. The 61 books are selected because different methods appearing in our experiments generates different true data. Thus it is more challenging than the first one. Similar to Luna Dong, we call it silver testing dataset. For both testing data, the true author list are manually assigned by people reading the cover page of the book. In our experiments, we will report the accuracy of each method on both testing dataset.

Since we do not have access to the pre-processed dataset used in previous works, we do the data cleaning by ourselves, and the clean data is posted online [2]. In the dataset, most stores separate the names by ";", but many others use ",". We manually recognize those stores and change them to names separated by ";". Then following procedures in [Dong et al., 2009], middle names are removed. Our dataset is cleaner compared to the data used in prior works, since, as we will see below, the voting results in our case is 82%, while past studies showed only 71%.

5.2.2 Baselines

There are 10 baseline models used in the experiments:

Voting. The candidate with max amount of providers is true data. If several candidates receive same voting, randomly pick one.

[1] http://lunadong.com/fusionDataSets.htm

[2] http://crystal.exp.sis.pitt.edu:8080/daz45/

Sums; Average.log; Investment; PooledInvestment; TruthFinder; Accuracy; AccuracySim. These seven discriminative methods have similar main idea, iterativly update each value's score and provider's reliability, only in different computing manners. First five methods appear in [Pasternack and Roth, 2010], and last two are proposed in [Dong and Srivastava, 2015]. TruthFinder and AccuracySim considers the similarity between candidate values, while other methods do not.

CATD; CRH. These two models are designed as generative model for numerical data, but can be adapted to categorical data as a discriminative model with slight modification. Each iteration, with evaluated provider's reliability, they try to generate/select estimated true value of each statement to minimize the difference between "estimated true matrix" and the observed input matrix [Li et al., 2014a,b, Zhao and Han, 2012]. Additionally, CATD is designed to smoothly predict truth on the long tail data with chi-squared distribution. The extra merits of first two methods is the lack of parameters.

5.2.3 Experiment settings

In this section, our proposed model is named as SO-Dis. It is based on our Subjective Opinion based data reliability assessment framework and predict true value in a discriminative manner.

Working on this categorical dataset, given statement (i.e., book) s_i and two value (i.e., book author list), we define the implication formula in Equation 26 and 28 as:

$$Imp(v_{ix} \rightarrow v_{iy}) = \frac{\#|v_{ix} \bigcap v_{iy}|}{\#|v_{iy}|}. \tag{36}$$

Following past studies, the parameters of all methods are set with optimal performance on the testing data. In TruthFinder, λ is set to be 0.4. In AccuracySim, λ is set to be 0.9. For the proposed method SOTD-Dis, both α and γ are set to be 0.2.

Table 8: Precision of eleven methods on true book author list finding task. Best results are in bold.

Method	Golden Testing	Silver Testing
SO-Dis	**0.94**	**0.776**
PooledInvestment	0.87	0.7275
TruthFinder	0.86	0.708
AccuracySim	0.91	0.689
Accuracy	0.89	0.689
Investment	0.79	0.634
Average.log	0.82	0.621
Voting	0.80	0.621
Sums	0.74	0.553
CRH	0.4	0.304
CATD	0.4	0.304

5.2.4 Experiment results

Precision of eleven methods are shown in Table 8, which is sorted by the performance on silver testing data. We can see that our proposed method SO-Dis has the best performance on both testing data. Further, SO-Dis increases precision by 3.3% compared with the second best method AccuracySim on the golden testing data; and is better than the second best method PooledInvestment by 6.7% on silver testing data. In addition, it seems that discriminative models have a much better performance than the CRH and CATD, which are modified to adapt this task. Also, methods (SO-Dis, TruthFinder, AccuracySim) that utilize the similarity/implication between values also shows a better performance than those who does not use.

5.3 Utility case 4: Find true city population

5.3.1 Dataset: Population

In this study, we pick the dataset Population, proposed in [Pasternack and Roth, 2010], to validate our proposed framework. This is a numerical dataset, a sample of Wikipedia edit history of city population. When the data was released in 2010, there were 44,761 tuples from 4,107 data providers. The version used in [Zhao and Han, 2012, Li et al., 2014b,a] contains 43,071 tuples. When we download it in 2019, it contains 51,761 tuples from 4,264 data providers on 40,583 cities. The testing data stays same, consisting of 308 randomly collected cities manually labeled with true population. Therefore, the experiment results differs from the results from past papers. We pre-process the dataset in a same way as [Zhao and Han, 2012, Li et al., 2014b,a]: (1) One provider may provide several population to same city, only the latest one is kept. (2) if a city only have one candidate value (from one or several providers), its data is removed. (3) Outliers on each city are removed in the same way as [Zhao and Han, 2012] with TruthFinder. After pre-processing, compared with 4,119 tuples on 1,148 cities from 2,415 providers are left and methods are evaluated on 274 cities [Zhao and Han, 2012, Li et al., 2014b,a], in our experiment dataset, 5,731 tuples on 1,814 cities from 2,467 providers are left, and methods are evaluated on 280 cities.

5.3.2 Baselines

There are 8 baseline models used in the experiments:

Voting. The candidate with max amount of providers is true data. If several candidates receive same voting, randomly pick one.

Median; Average. The median and average of all candidate values is predicted as true.

Investment; TruthFinder. These two discriminative methods have similar main idea, iterativly update each value's score and provider's reliability, only in different computing manners. Investment in [Pasternack and Roth, 2010] doesn't consider the value similarity among candidate values, while TruthFinder in [Dong and Srivastava, 2015] considers it.

CATD; CRH; GTM. These three models are designed as generative model for numerical data, but can be adapted to categorical data as a discriminative model with slight modification. Each iteration, with evaluated provider's reliability, they try to generate/select estimated true value of each statement to minimize the difference between "estimated true matrix" and the observed input matrix [Li et al., 2014a,b, Zhao and Han, 2012]. Additionally, CATD is designed to smoothly predict truth on the long tail data with chi-squared distribution. The extra merits of first two methods is the lack of parameters.

5.3.3 Experiment Settings

In this section, we will test two proposed model: SO-Dis and SO-Gen. The former one is model based on our Subjective Opinion based data reliability assessment framework and predict true value in a discriminative manner, and the latter one is model predicting true value in a generative manner.

Working on this numerical dataset, given statement (i.e., city) s_i, we define the implication formula in Equation 26 and 28 as:

$$Imp(v_{ix} \to v_{iy}) = 1 - \frac{|v_{ix} - v_{iy}|}{max(\{v_{ij}|j \in \{1, \dots, m\}) - min(\{v_{ij}|j \in \{1, \dots, m\})}. \tag{37}$$

Following [Zhao and Han, 2012, Li et al., 2014b,a], three evaluation metrics are selected: MAE, RMSE, and Error Rate. In terms of Error Rate, "error" appears when the predicted truth is smaller or larger than the ground truth by 10%.

Similarly, following past studies, the parameters of all methods are set based on optimal performance on the testing data. In TruthFinder, λ is set to be 0.3. In terms of GTM, we have two set of parameters, the first being ($\alpha = 10, \beta = 10, \mu_0 = 0, \sigma_0^2 = 1$) suggested by Zhao and Han [2012], and the second being ($\alpha = 4, \beta = 1, \mu_0 = 0, \sigma_0^2 = 1$), which has best performance in our experiment. For the proposed method SO-Dis, α is set to be 0.1. Finally, for SO-Gen, α is set to be 0.01, and β is set to be 0.01.

Table 9: Precision of eleven methods on true book author list finding task. First group shows the performance of six discriminative models, and second group shows the performance of six generative models. Best results are in bold.

	Methods	**MAE**	**RMSE**	**Error Rate**
Dirscriminative models	SO-Dis	**1122.71**	**4845.73**	**14.0%**
	TruthFinder	1744.05	8942.86	17.0%
	Voting	2462.28	11350.62	22.8%
	Investment	2614.21	11378.42	26.0%
	CRH-weighted median	3030.23	12696.96	26.0%
	Median	2426.17	9753.68	33.5%
Generative models	SO-Gen	**1511.47**	**7211.25**	27.9%
	CATD	1796.67	8765.81	**21.3%**
	GTM-parameter by us	2424.10	8659.36	57.0%
	GTM-parameter in [Zhao and Han, 2012]	2710.30	9290.32	58.1%
	Average	3231.97	9768.31	57.4%
	CRH-weighted average	3805.10	11898.04	58.5%

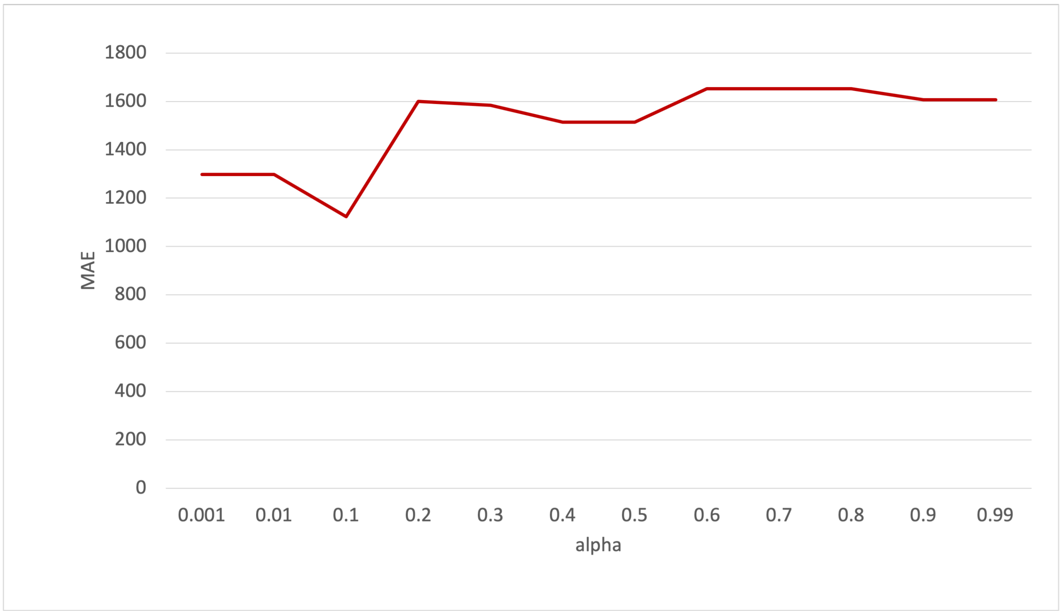
5.3.4 Experiment Results

All methods' performance is shown in Table 9. We can see that the proposed method SO-Dis gives best performance on all three metrics. Additionally, SO-Gen gives the second best on MAE and RMSE, while CATD gives second best Error Rate. Compared with first group (discriminative models), second group (generative models) generally have a relatively similar RMSE but and a higher Error Rate. It indicates that the "error cases" in first group, though has the less quantity, are farther away from truth than that of second group. Apart from our two proposed models, CATD and TruthFinder are the baseline models giving best performance.

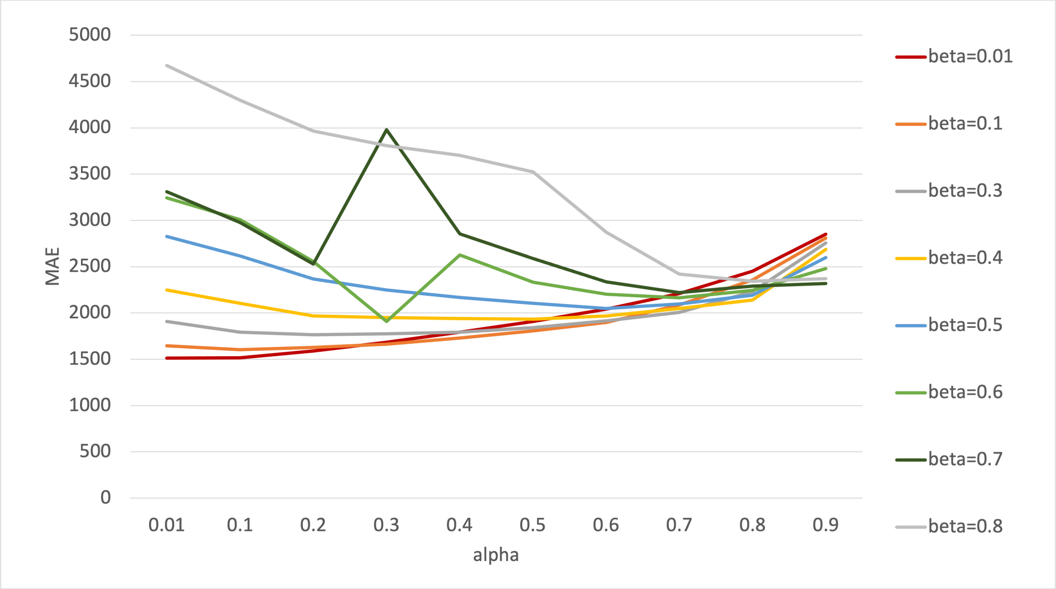
5.3.5 Further Discussion

Our proposed SO-Dis model has a parameter, α, describing model's basic uncertainty towards each provider. SO-Gen has one more parameter, β, describing provider's basic uncertainty towards each value. In the above experiment results subsection, following baseline paper, these parameters and all baseline parameters are all selected to achieve the highest performance with the Population dataset. We report the model's MAE performance with different parameters in Figrure 21.

Figure 21(a) shows SO-Dis's MAE performance with different α. We could see that SO-Dis has a relatively better performance when α is small. However, even though α is large, SO-Dis MAE is still much less than TruthFinder and CATD (baseline best performance). Therefore, we could say SO-Dis is robust to parameter α. Figure 21(b) shows SO-Gen's MAE performance with different α and β. We can find that SO-Gen is more sensitive to the parameter settings. Parameter α describes model's basic uncertainty towards each provider, while β describes provider's basic uncertainty towards the temporary true value $v_{i'}$. SO-Gen has the best performance when the model have lowest uncertainty ($\{\alpha, \beta\}$) towards the data. If model keeps low uncertainty (α) towards providers but provider increases uncertainty (β) towards the values, performance drops a lot. If keep provider's uncertainty (β) towards the values and increase model's uncertainty (α) towards providers, we could also see performance drops, but not so largely. We could say SO-Gen is sensitive to β, and less sensitive to α.



(a) SO-Dis model's MAE sensitivity to α (alpha).



(b) SO-Gen model's MAE sensitivity to α (alpha) and β (beta).

Figure 21: MAE of SO-Dis and SO-Gen varying with different $\{\alpha, \beta\}$ on Population dataset.

5.4 Summary

In this chapter, we have presented the Subjective Opinion based data reliability assessment framework in the scenario statement candidate value conflicts appear, which provide rich evidences for model to accurately make assessment. There is no labeled training data in this scenario, and hence is moderate challenging. We validate our framework on two real-world datasets. Experiment results shows that our proposed models significantly outperforms state-of-art baseline models, validating the effectiveness of our proposed framework.

6.0 RQ 4: How to assess data reliability with historical data when sources provide conflicting values

In this chapter, we will discuss the data reliability assessment in the scenario where historical data are provided, and also for a given statement, the conflicting values appear. This question is supposed to be the easiest one in the four research questions, as both the historical data and conflicting values could provide additional evidences for reliability assessment.

6.1 Subjective Opinion based data reliability assessment

Subjective Opinion based data reliability assessment framework for this scenario is quite similar to the one in Scenario 3. Provider reliability is represented by $\omega_{p_j}^{model}$, describing the model's opinion towards the declaration "provider p_j is reliable". One statement may receive several different values from providers, and hence we evaluate the reliability of candidate value v instead of statement s_i, which is represented by ω_{v}^{model}. In such a scenario, the proposed framework considers all the assumptions talked in Scenario 3. My framework first initialize the statement true values, then iteratively assess provider/source reliability and update the statement true value until converge, as shown in Figure 22. In this scenario, labeled training data (i.e., statement true values) is provided. Therefore, we could directly use the provided true values in the iteration and expect the reliability assessment is more accurate than that in Scenario 3.

Step-1: initialize the statement true values and statement discrimination score. If the statement s_i has the ground-truth value $v_{i'}$, this true value is directly used. For those without labels, select true value $v_{i'}$ based on naive strategies, such as Majority Voting, Average, Maximum, Minimum, or other algorithms.

In Scenario 3, each statement has a discrimination score, describing referencing the statement's data, model's ability to differentiate the reliable and unreliable providers. It was

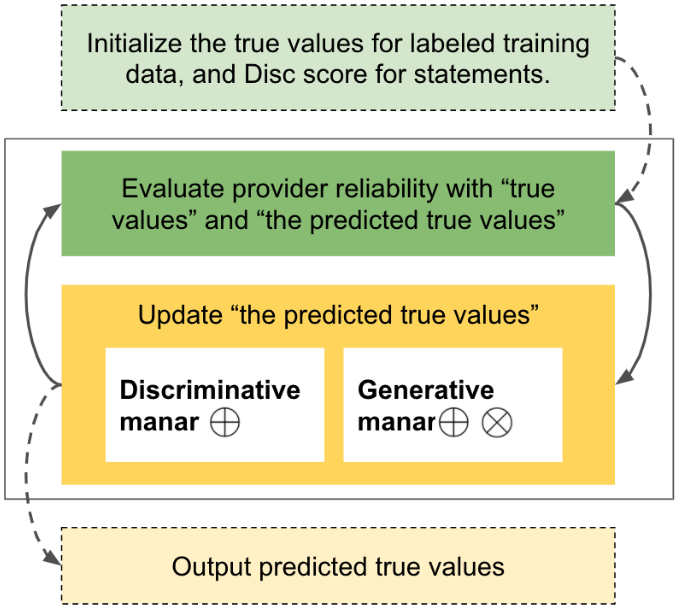

Initialize the true values for labeled training data, and Disc score for statements.
Evaluate provider reliability with “true values” and “the predicted true values”
Update “the predicted true values”
Discriminative manar ⊕
Generative manar⊕ ⊗
Output predicted true values

Figure 22: Subjective Opinion based data reliability assessment framework for Scenario 4.

calculated with value distribution in RQ3, and in this scenario, we also pick this concept but re-define its formula: if the statement has the labeled true value, then it receives a full discrimination score, $Disc(s_i) = 1$; otherwise, the discrimination score is $\gamma \in (0, 1)$.

Step-2: calculate provider's reliability. Provider's reliability is defined as our opinions towards the statement "provider p_j is reliable", which is define as:

$$\omega^{model}_{p_j} = \{t^{model}_{p_j}, d^{model}_{p_j}, u^{model}_{p_j}\}. \tag{38}$$

Intuitively, if the provider's values are close to the statement true values, then this provider is reliable; otherwise, this provider is unreliable. Therefore, we define:

$$\begin{cases} t^{model}_{p_j} = \frac{\sum_{i=1}^{n} Disc(s_i)*Imp(v_{i'}\rightarrow v_{ij})}{\sum_{i=1}^{n}\{Disc(s_i)|v_{ij}\neq null\}} * (1 - u^{model}_{p_j}) \\ d^{model}_{p_j} = 1 - t^{model}_{p_j} - u^{model}_{p_j} \\ u^{model}_{p_j} = \theta * \frac{\sum_{i=1}^{n}\{1-Disc(s_i)|v_{ij}\neq null\}}{\sum_{i=1}^{n}\{1|v_{ij}\neq null\}} + \alpha, \end{cases} \tag{39}$$

where α is a numeric value, representing our basic uncertainty, θ controls the effect of providng values for low discriminative statements, and $Imp(v_{i'} \rightarrow v_{ij})$ is the implication from ground-truth/predicted true value to v_{ij}, which is defined based on specific task background.

Step-3: update the true values. If the statement has the labeled true value, the true value is directly used. For those without labels, true value could be decided based on each value's reliability, either in a discriminative way, or in a generative way. If values are numeric, we could select the value with highest reliability as the true value, and could also calculate a true value based on the reliability score. If values are categorical, then we will have to chose one from the existing candidate values.

Case 1: in a discriminative manner. If provider/source p_j has same belief in all values that he provides, then directly fuse these sources' reliability with Subjective Logic consensus operation. The fused opinion represents our belief for the statement "v_{ij} is the true value for statement s_i", which is represented as:

$$\omega_{v_{ij}}^{model} = \{t_{v_{ij}}^{model}, d_{v_{ij}}^{model}, u_{v_{ij}}^{model}\}. \tag{40}$$

It is obtained by:

$$\omega_{v_{ij}}^{model} = \omega_{p_1}^{model} \oplus \omega_{p_2}^{model} \oplus ... \oplus \omega_{p_k}^{model}, \tag{41}$$

where p_1, p_2,..., p_k are providers who provide values for s_i. Given a statement s_i, for each candidate value $\{v_{ij} | j = 1, ..., m,\ v_{ij} \neq null\}$, we compare their reliability $\omega_{v_{ij}}^{model}$, and select the value with highest trust $t_{v_{ij}}^{model}$ as the true value.

Case 2: in a generative manner. In this case, for the statement s_i without ground-truth label, we assume that value $v_{i''}$ is the temporary true value. Then provider p_j gives an opinion towards "the value $v_{i''}$ is the true value for statement s_i" based on the distance between his provided value v_{ij} and this $v_{i''}$. The model could learn $v_{i''}$'s reliability based on p_j's recommendation. After considering statement s_i all providers' recommendation, the model could get a fused idea about this temporary true value $v_{i''}$, and based on which generates a predicted true value.

There are many ways to obtain the temporary true value $v_{i''}$, and we recommend to use statement's maximum or minimum candidate value. Take maximum for example, after obtaining the final opinion towards this $v_{i''}$, if trust is high, the generated value should approach the maximum candidate value; if trust is low, the generated value should approach the minimum candidate value. Vice versa. On the other hand, if chose median, average, or majority, when the trust in the final fused opinion is low, the model has no idea about whether the predicted true value should be bigger or lower than $v_{i''}$. In this study, for each statement s_i that has no label, we chose its maximum candidate value as the temporary true value, $v_{i''} = max(\{v_{ij}|j \in \{1, \ldots, m\})$.

First, on each statement s_i, we normalize all the candidate values in the following manner:

$$v_{ij}^{'} = \frac{v_{ij} - min(\{v_{ij}|j \in \{1, \ldots, m\})}{max(\{v_{ij}|j \in \{1, \ldots, m\}) - min(\{v_{ij}|j \in \{1, \ldots, m\})}, \tag{42}$$

Then we have $v_{ij}^{'} \in [0, 1]$. After this, statement "true value of statement s_i is the max candidate value $max(\{v_{ij}|j \in \{1, \ldots, m\})$" is mapped to "in the normalized space, true value of statement s_i is 1". Thereby, the provider p_j's opinion towards the statement can be defined as:

$$\omega_{v_{i''}=1}^{p_j} = \{(1-\beta)v_{ij}^{'}, 1-(1-\beta)v_{ij}^{'} - \beta, \beta\}, \tag{43}$$

where β describe provider's fundamental uncertainty, similar to above α.

Second, the provider p_j can recommend his opinion $\omega_{v_{i''}=1}^{p_j}$ to the model. Recommendation operation can help people know the statement according to their acquaintances. Thus, model's opinion towards "in the normalized space, true value of statement s_i is 1" could be obtained:

$$\omega_{v_{i''}=1}^{model,p_j} = \omega_{p_j}^{model} \otimes \omega_{v_{i''}=1}^{p_j}. \tag{44}$$

Statement s_i has a set of candidate values from several providers $\{p_1, p_2, ..., p_k\}$, and the model should have a summarized opinion based on all recommendations. Consensus operation can help fuse several opinions towards one statement together. The model's final opinion towards the temporary true value $v_{i''}$ is defined as:

$$\omega_{v_{i''}=1}^{model,p_1,p_2,\ldots,p_k} = \omega_{v_{i''}=1}^{model,p_1} \oplus \omega_{v_{i''}=1}^{model,p_2} \oplus \ldots \oplus \omega_{v_{i''}=1}^{model,p_k}. \tag{45}$$

In the final fused opinion $\omega_{v_{i''}=1}^{model,p_1,p_2,\ldots,p_k}$, the trust reflects model's confidence about the temporary true value $v_{i''} = 1$ in the normalized space. There are two ways to get the final predicted true value:

Unsupervised generation. Map $t_{v_{i''}=1}^{model,p_1,p_2,\ldots,p_k}$ to the original numerical space, and use it as the predicted true value $v_{i'}$.

$$\begin{aligned} v_{i'} =& t_{v_{i''}=1}^{model,p_1,p_2,\ldots,p_k} * (max(\{v_{ij}|j \in \{1,\ldots,m\}) - min(\{v_{ij}|j \in \{1,\ldots,m\}))+ \\ & min(\{v_{ij}|j \in \{1,\ldots,m\}). \end{aligned} \tag{46}$$

Supervised generation. Since we have labeled training data, we could build a regression model to predict the true value $v_{i'}$. The $\{t_{v_{i''}=1}^{model,p_1,p_2,\ldots,p_k}, max(\{v_{ij}|j \in \{1,\ldots,m\}),$ $min(\{v_{ij}|j \in \{1,\ldots,m\})\}$ and their transformed versions will be the features.

6.2 Utility case 6: Find true city population

6.2.1 Dataset: Population

In this study, we pick the dataset Population, proposed in [Pasternack and Roth, 2010], to validate our proposed framework. This is a numerical dataset, a sample of Wikipedia edit history of city population. When the data was released in 2010, there were 44,761 tuples from 4,107 data providers. The version used in [Zhao and Han, 2012, Li et al., 2014b,a] contains 43,071 tuples. When we download it in 2019, it contains 51,761 tuples from 4,264 data providers on 40,583 cities. It originally contains ground-truth populations only for 274 cities, but in this study, we need ground-truth labels for all cities. Therefore, we label the Population dataset. First, we download the United States 2000 Census data. Then, we scan the whole city set in Population and Census data, and all cities are indexed by "city name, state name". Only the cities that has one and only one exact match in both data are kept. Also, these cities' related wiki editors are kept.

Then, I pre-process the dataset in a same way as [Zhao and Han, 2012, Li et al., 2014b,a]: (1) One provider may provide several population to same city, only the latest one is kept. (2) if a city only have one candidate value (from one or several providers), its data is removed. (3) Outliers on each city are removed in the same way as Zhao and Han [2012] did with TruthFinder. In this way, we obtain 13,359 tuples from 667 providers about 10,970 cities.

6.2.2 Baselines

There are 7 baseline models in this section.

Voting. The candidate with max amount of providers is true data. If several candidates receive same voting, randomly pick one.

Median; Average. The median and average of all candidate values is predicted as true.

CRH; GTM. Model CRH Li et al. [2014b] is designed as generative model for numerical data, but can be used as a discriminative model with slight modification. Model GTM Zhao and Han [2012] is a generative model, and we did not further extend it as the discriminative model in this study. Each iteration, with evaluated provider's reliability, they try to generate/select estimated true value of each statement to minimize the difference between "estimated true matrix" and the observed input matrix . They are unsupervised models, and we did a small modification to adapt them into a semi-supervised mode: when CRH and GTM update the "estimated true values" in each iteration, they skip those statements (i.e., city) having ground-truth values (i.e., populations).

SSTF. Model SSTF [Yin and Tan, 2011] is a semi-supervised model. It also runs iteratively updating the city's population values, and skip those statement with ground-truth values. SSTF maintains two weight matrix, one being weights for values from same sources (i.e., wiki editor), and one being weights for values on same statement (i.e., city). SSTF believes that values of same sources or on same statements should be similar, and lower down the weights of those having big differences. It tries to find the optimal weight matrix, and then select the best value that are weighted closest to all other values.

OpSTD. Model OpSTD Yang et al. [2018] is a semi-supervised model. It iteratively update providers' reliability and the expected values. Similarly, it ignore the statements (i.e., city) having ground-truth values (i.e., populations). Expected values are calculated as the weighted average values, where weight is provider's weight.

6.2.3 Experiment Settings

Following [Zhao and Han, 2012, Li et al., 2014b,a], three evaluation metrics are selected: MAE, RMSE, and Error Rate. In terms of Error Rate, "error" appears when the predicted truth is smaller or larger than the ground truth by 10%.

Similarly, following past studies, the parameters of all methods are set based on optimal performance on the testing data. In TruthFinder, λ is set to be 0.3. In terms of GTM, we have two set of parameters, the first being ($\alpha = 10, \beta = 10, \mu_0 = 0, \sigma_0^2 = 1$) suggested by Zhao and Han [2012], and the second being ($\alpha = 4, \beta = 1, \mu_0 = 0, \sigma_0^2 = 1$), which has best performance in our experiment. For our proposed method SO-Dis, γ is set to be 0, while α is set to be 0.01. Finally, for SO-Gen, γ is set to be 0, while α is set to be 0.01, and β is set to be 0.01.

6.2.4 Experiment Results

All methods' performance is shown in Table 10. We can see that the proposed method SO-Gen gives best performance on all three metrics. Additionally, SO-Dis gives the second best on all three metrics. Actually, all models, except SO-Gen, give similar error rate without significant differences (p-value¿0.05).

We can also see following findings: (1) It is reasonable to use predictions from TruthFinder as priors to remove outliers, consistent with findings from [Zhao and Han, 2012]. Naive methods, especially Average, gives a much worse performance. (2) Second group (generative models) usually have a relatively smaller RMSE and and a higher Error Rate than the third group (discriminative models), indicating that either the "correct cases" whose distance is smaller than 10% from truth in the second group are more accurate than that in the third group, or the "error cases" in third group are farther from truth than that of second group. (3) Also, the lower Error Rate in third group means that true value usually appears in the candidate value set. Also, we tried to run TruthFinder after outlier removed, but it does not provide further improvement, and even declined a little bit in terms of MAE and RMSE.

6.2.5 Further Discussion

Our proposed SO-Dis model has two parameters, γ for discriminative score of statement without labels and α for model's basic uncertainty towards each provider. SO-Gen has one more parameter, β, describing provider's basic uncertainty towards each value. In the above

Table 10: Compare ten models' performance on Population. First group shows the performance of discriminative models; second group models predict truth in a generative way. Best results in both groups are in bold.

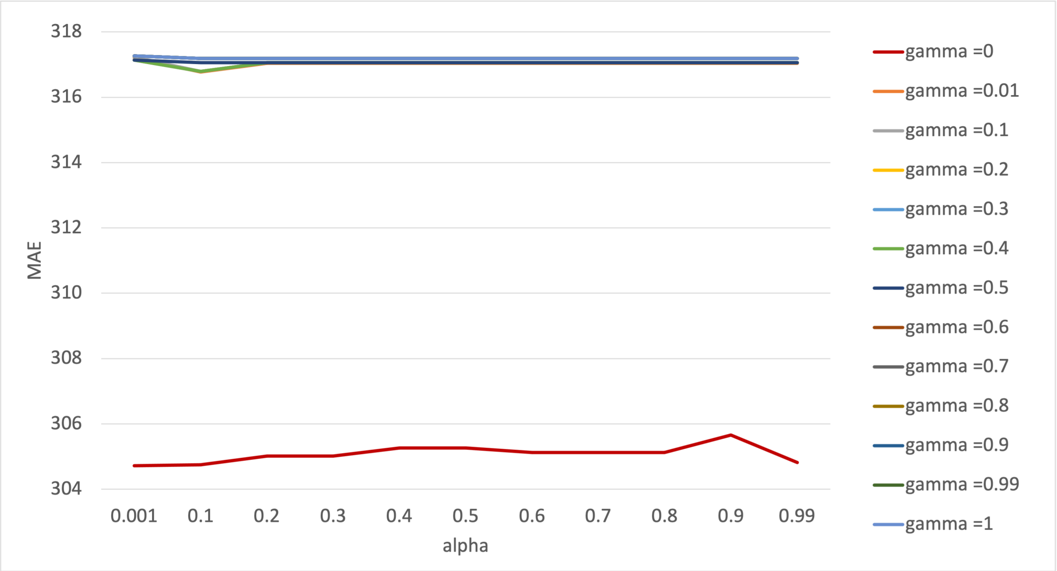
	Models	**MAE**	**RMSE**	**Error rate**
Discriminative models	**SO-Dis**	**304.718**	**7468.288**	0.816
	median	310.489	13673.871	0.817
	voting	314.674	13771.172	0.816
	CRH-selective	351.850	8377.411	0.818
	SSTF	717.552	12514.518	0.819
Generative models	**SO-Gen**	**190.416**	**3350.384**	**0.481**
	GTM	311.389	7509.255	0.818
	OpSTD	337.331	7659.214	0.819
	average	373.700	13792.081	0.819
	CRH-generative	409.939	8770.881	0.819

experiment results subsection, these parameters and all baseline parameters are all selected to achieve the highest performance with the Population dataset in cross-validation. We repeat the cross-validation procedure 30 times, and the average performance is reported in Figure 23.

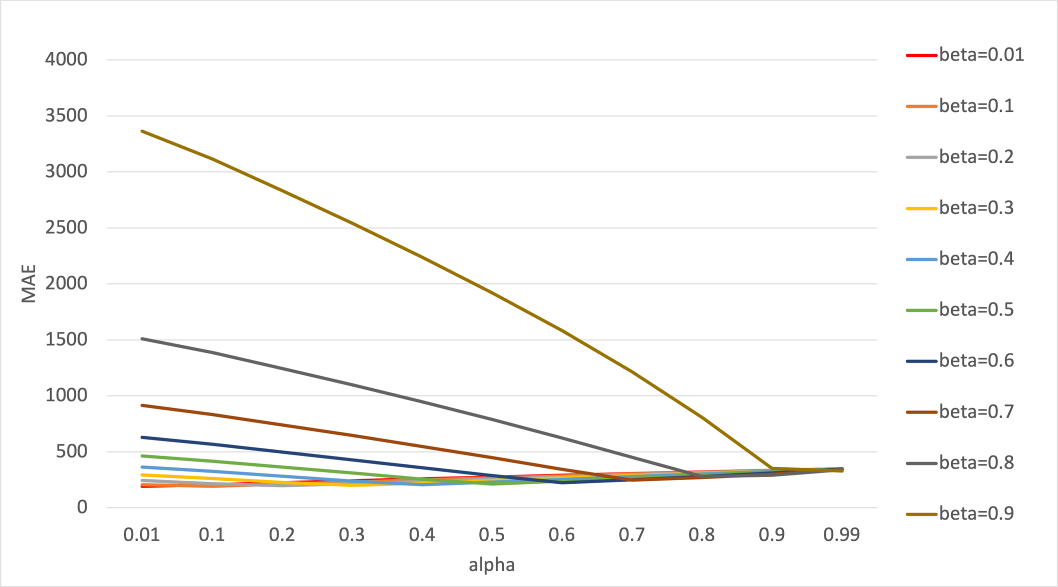
Figure 23(a) shows SO-Dis's MAE performance with different α and γ. Surprisingly, SO-Dis obtains best performance when $\gamma = 0$. It indicates that the provider's reliability is totally decided by the statements with ground-truth labels. If they provides approaching truth value, then provider has higher reliability; otherwise, provider will have low reliability. One reason maybe that current experiments are with five-fold cross-validation, i.e., 80% statements are equipped with ground-truth value in the iterations. If we lower down the size of labeled training data, reliability assessment then have to rely on unlabeled statement, and γ is expected to increase.

On the other hand, when parameter γ is fixed, α's variation slightly affect MAE performance. But still, SO-Dis gets best performance when $\alpha = 0.001$. It indicates that model have very low uncertainty about the Population dataset. One reason is above mentioned five-fold cross-validation, with 80% statements having labels, and hence uncertainty is low. Another reason is that the conflicting values also provide the model with reliability evidences, further lowering down model's uncertainty.

Figure 23(b) shows SO-Gen's MAE performance with different α and β, while γ is set as 0. In our experiments, I find SO-Gen has a same trend as SO-Dis on $gamma$, hence directly use $gamma = 0$ and only illustrate SO-Gen's MAE sensitivity with α and β. Parameter α describes model's basic uncertainty towards each provider, while β describes provider's basic uncertainty towards the temporary true value $v_{i'}$. SO-Gen has the best performance with lowest $\{\alpha, \beta, \gamma\}$. From Figure 23(b) we could see that when β is very small, such as $\beta \in [0.01\ 0.3]$, SO-Dis's performance is quite robust to parameter α. It indicates that, as long as SO-Dis's providers are confident (low uncertainty with β) about their judgement of statement temporary true value $v_{i'}$, the final prediction of the truth will have good quality. However, if providers have high uncertainty (β) about the statement temporary true value $v_{i'}$, but model have low uncertainty (α) with providers, final prediction is far away from the real truth, leading poor MAE performance.



(a) SO-Dis model's MAE sensitivity to α (alpha) and γ (gamma).



(b) SO-Gen model's MAE sensitivity to α (alpha) and β (beta) with γ set as 0.

Figure 23: MAE of SO-Dis and SO-Gen varying with different $\{\gamma, \alpha, \beta\}$ on Population dataset.

To summary, Scenario 4 is the most certain scenario in four research questions, and the Population datasets validate this funding as two proposed models get best performance when all uncertainty related parameters $\{\alpha, \beta, \gamma\}$ is assigned with the lowest value.

6.3 Summary

In this chapter, we have presented the Subjective Opinion based data reliability assessment framework in the scenario where statement historical labels are provided, and statement candidate value conflicts appear, both of which provide rich evidences for model to accurately make assessment. This scenario is the most easy one in this whole dissertation study. We also validate our framework on a real-world dataset. Experiment results shows that our proposed models significantly outperforms state-of-art baseline models, validating the effectiveness of our proposed framework.

7.0 Conclusion

Data reliability has always been considered important, especially in current society, where real and fake data from diverse sources fluctuates people's daily life. In this dissertation, I work on accurate data reliability assessment to help people get good quality data. I found past models do not fully consider the uncertainty in the dataset, and hence in this work, I propose series of data reliability assessment frameworks for different scenarios. Also, I found Subjective Opinion is naturally good at recording data uncertainty and introducing it into my frameworks. Experiments on multiple real-world datasets show the effectiveness of the proposed models and the framework. In this dissertation, I make the following key contributions:

- In Chapter 3, I identify a data reliability assessment scenario where historical labels are not available for training and statement candidate values have no conflicts. Data reliability assessment in this scenario is more challenging than in other scenarios. It is hard to collect evidence to differentiate between reliable or unreliable data. In this scenario, the model has to be designed with background knowledge. Therefore, first, I build a general framework and then give specific designs with a real-world problem, finding cancer driver genes with TCGA data. A lot of gene mutation background knowledge is used to build the data reliability model. I conduct experiments on TCGA data, and my model outperforms state-of-art baseline models, validating the proposed framework's effectiveness.
- In Chapter 4, I identify a data reliability assessment scenario where historical labels are available for training (statement candidate values have no conflicts). Data reliability assessment in this scenario is moderately challenging. The labeled training data can be used to differentiate between reliable or unreliable data. Existing models do not fully consider the uncertainty in the data. Therefore, I propose a new data reliability assessment framework, which is based on Subjective Opinion and can handle data uncertainties well. People could adapt it to specific tasks with minor changes. I conduct experiments on a fake news detection task with a real-world dataset. My model outperforms state-of-art

baseline models, validating the effectiveness of the proposed framework.

- In Chapter 5, I identify a data reliability assessment where though historical labels are not available for training, each statement has one or candidate values, whose distribution provides evidence to differentiate reliable or unreliable data. Data reliability assessment in this scenario is moderately challenging. My proposed Subjective Opinion based framework is able to comprehensively handle the data uncertainty and also introduce the statement Discrimination Score to describe the statement's ability to provide evidence for the model, which is new in this area. I conduct experiments on two real-world datasets, and my models outperform state-of-art baseline models, validating my proposed framework's effectiveness.
- In Chapter 6, I identify a data reliability assessment scenario where historical labels are provided and also each statement has one or candidate values, whose distribution provides evidence to differentiate reliable or unreliable data. Data reliability assessment in this scenario is easier than in other scenarios. My proposed Subjective Opinion based framework is able to comprehensively handle the data uncertainty and also introduce the statement Discrimination Score to describe the statement's ability to provide evidence for the model. I constructed a labeled real-world dataset, and validates that my model outperform state-of-art baseline models.

In this study, I find in these different scenarios, if real data dominates the dataset, my proposed model and listed baseline models usually could gain good performance. However, when misinformation and uncertain data fluctuates the dataset, all models' performance will drop. Also, according to the experiments on TCGA dataset in Section 3.2, we could find that domain knowledge is very informative, providing much evidences to differentiate the reliable and unreliable data. Currently, we equally regard each data provider, whose reliability is decided by the quality of their provide values. In the future, I will consider the data provider's background into the reliability assessment framework. I.E., maybe domain experts will have higher reliability in his/her domain, and lower reliability in other domains.

Bibliography

O. Ajao, D. Bhowmik, and S. Zargari. Fake news identification on twitter with hybrid cnn and rnn models. In *Proceedings of the 9th International Conference on Social Media and Society*, pages 226–230, 2018.

S. Banerji, K. Cibulskis, C. Rangel-Escareno, K. K. Brown, S. L. Carter, A. M. Frederick, M. S. Lawrence, A. Y. Sivachenko, C. Sougnez, L. Zou, et al. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, 486(7403):405–409, 2012.

R. E. Barlow and F. Proschan. Statistical theory of reliability and life testing: probability models. Technical report, Florida State Univ Tallahassee, 1975.

A. Bashashati, G. Haffari, J. Ding, G. Ha, K. Lui, J. Rosner, D. G. Huntsman, C. Caldas, S. A. Aparicio, and S. P. Shah. Drivernet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome biology*, 13(12):R124, 2012.

T. R. Bennett, J. M. Booker, S. Keller-McNulty, and N. D. Singpurwalla. Testing the untestable: Reliability in the 21st century. *IEEE Transactions on Reliability*, 52(1):118–124, 2003.

L. Bertossi. Consistent query answering in databases. *ACM Sigmod Record*, 35(2):68–76, 2006.

L. Bertossi and J. Chomicki. Query answering in inconsistent databases. In *Logics for emerging applications of databases*, pages 43–83. Springer, 2004.

J. Bleiholder and F. Naumann. Data fusion. *ACM Computing Surveys (CSUR)*, 41(1):1, 2009.

A. Bondielli and F. Marcelloni. A survey on fake news and rumour detection techniques. *Information Sciences*, 497:38–55, 2019.

L. Cartegni, S. L. Chew, and A. R. Krainer. Listening to silence and understanding nonsense: exonic mutations that affect splicing. *Nature reviews genetics*, 3(4):285–298, 2002.

H. Carter, S. Chen, L. Isik, S. Tyekucheva, V. E. Velculescu, K. W. Kinzler, B. Vogelstein, and R. Karchin. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer research*, 69(16):6660–6667, 2009.

C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684, 2011.

Y. Chen, J. Hao, W. Jiang, T. He, X. Zhang, T. Jiang, and R. Jiang. Identifying potential cancer driver genes by genomic data integration. *Scientific reports*, 3:3538, 2013.

Y. Chen, N. J. Conroy, and V. L. Rubin. Misleading online content: Recognizing clickbait as false news. In *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*, pages 15–19. ACM, 2015.

Y. Chen, L. Chen, and C. J. Zhang. Crowdfusion: A crowdsourced approach on data fusion refinement. In *Data Engineering (ICDE), 2017 IEEE 33rd International Conference on*, pages 127–130. IEEE, 2017.

C.-C. Chiu, Y.-S. Yeh, and J.-S. Chou. An effective algorithm for optimal k-terminal reliability of distributed systems. *Malaysian Journal of Library & Information Science*, 6(2): 101–118, 2001.

Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference*, pages 21–30. ACM, 2010.

G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193, 2015.

G. Ciriello, E. Cerami, C. Sander, and N. Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome research*, 22(2):398–406, 2012.

F. S. Collins and V. A. McKusick. Implications of the human genome project for medical science. *Jama*, 285(5):540–544, 2001.

N. J. Conroy, V. L. Rubin, and Y. Chen. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, 52(1):1–4, 2015.

N. D. Dees, Q. Zhang, C. Kandoth, M. C. Wendl, W. Schierding, D. C. Koboldt, T. B. Mooney, M. B. Callaway, D. Dooling, E. R. Mardis, et al. Music: identifying mutational significance in cancer genomes. *Genome research*, 22(8):1589–1598, 2012.

M. L. Della Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro, and L. de Alfaro. Automatic online fake news detection combining content and social signals. In *2018 22nd Conference of Open Innovations Association (FRUCT)*, pages 272–279. IEEE, 2018.

S. Destercke, P. Buche, and B. Charnomordic. Data reliability assessment in a data warehouse opened on the web. In *International Conference on Flexible Query Answering Systems*, pages 174–185. Springer, 2011.

L. Ding, G. Getz, D. A. Wheeler, E. R. Mardis, M. D. McLellan, K. Cibulskis, C. Sougnez, H. Greulich, D. M. Muzny, M. B. Morgan, et al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 455(7216):1069–1075, 2008.

X. L. Dong and F. Naumann. Data fusion: resolving data conflicts for integration. *Proceedings of the VLDB Endowment*, 2(2):1654–1655, 2009.

X. L. Dong and D. Srivastava. Big data integration. *Synthesis Lectures on Data Management*, 7(1):1–198, 2015.

X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. *Proceedings of the VLDB Endowment*, 2(1):550–561, 2009.

X. L. Dong, B. Saha, and D. Srivastava. Less is more: Selecting sources wisely for integration. In *Proceedings of the VLDB Endowment*, volume 6, pages 37–48. VLDB Endowment, 2012.

X. L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang. Knowledge-based trust: Estimating the trustworthiness of web sources. *Proceedings of the VLDB Endowment*, 8(9):938–949, 2015.

X. S. Fang, Q. Z. Sheng, X. Wang, and A. H. Ngu. Smartmtd: A graph-based approach for effective multi-truth discovery. *arXiv preprint arXiv:1708.02018*, 2017a.

X. S. Fang, Q. Z. Sheng, X. Wang, and A. H. Ngu. Value veracity estimation for multi-truth objects via a graph-based approach. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 777–778. International World Wide Web Conferences Steering Committee, 2017b.

X. S. Fang, Q. Z. Sheng, X. Wang, W. E. Zhang, and A. H. Ngu. From appearance to essence: Comparing truth discovery methods without using ground truth. *arXiv preprint arXiv:1708.02029*, 2017c.

P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton. A census of human cancer genes. *Nature reviews cancer*, 4(3):177–183, 2004.

A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 131–140. ACM, 2010.

A. Gonzalez-Perez and N. Lopez-Bigas. Functional impact bias reveals cancer drivers. *Nucleic acids research*, 40(21):e169–e169, 2012.

M. Götzinger, A. Anzanpour, I. Azimi, N. Taherinejad, and A. M. Rahmani. Enhancing the self-aware early warning score system through fuzzified data reliability assessment. In *International Conference on Wireless Mobile Communication and Healthcare*, pages 3–11. Springer, 2017.

C. Greenman, P. Stephens, R. Smith, G. L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague, A. Butler, C. Stevens, et al. Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132):153–158, 2007.

G. B. Guacho, S. Abdali, N. Shah, and E. E. Papalexakis. Semi-supervised content-based detection of misinformation via tensor embeddings. In *2018 IEEE/ACM International*

Conference on Advances in Social Networks Analysis and Mining (ASONAM), pages 322–325. IEEE, 2018.

M. Gupta, P. Zhao, and J. Han. Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 153–164. SIAM, 2012.

G. Hardy, C. Lucet, and N. Limnios. K-terminal network reliability measures with binary decision diagrams. *IEEE Transactions on Reliability*, 56(3):506–515, 2007.

N. Hassan, C. Li, and M. Tremayne. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th acm international on conference on information and knowledge management*, pages 1835–1838. ACM, 2015.

J. P. Hou and J. Ma. Dawnrank: discovering personalized driver genes in cancer. *Genome medicine*, 6(7):56, 2014.

C. Jaggle, J. Neidig, T. Grosch, and F. Dressler. Introduction to model-based reliability evaluation of wireless sensor networks. In *2nd IFAC workshop on dependable control of discrete systems*, pages 149–154. Citeseer, 2009.

Z. Jin, J. Cao, Y. Zhang, and J. Luo. News verification by exploiting conflicting social viewpoints in microblogs. In *Thirtieth AAAI conference on artificial intelligence*, 2016.

E. J. Jordan and R. Radhakrishnan. Machine learning predictions of cancer driver mutations. In *Proceedings of the 2014 6th International Advanced Research Workshop on In Silico Oncology and Cancer Investigation-The CHIC Project Workshop (IARWISOCI)*, pages 1–4. IEEE, 2014.

A. Jøsang. Artificial reasoning with subjective logic. In *Proceedings of the second Australian workshop on commonsense reasoning*, volume 48, page 34. Citeseer, 1997.

A. Josang. Conditional reasoning with subjective logic. *Journal of Multiple-Valued Logic and Soft Computing*, 15(1):5–38, 2008.

A. Jøsang. *Subjective logic*. Springer, 2016a.

A. Jøsang. Opinion representations. In *Subjective Logic*, pages 19–50. Springer, 2016b.

A. Jøsang. *Subjective logic*. Springer, 2016c.

A. Jøsang, R. Hayward, and S. Pope. Trust network analysis with subjective logic. In *Proceedings of the 29th Australasian Computer Science Conference-Volume 48*, pages 85–94. Australian Computer Society, Inc., 2006.

K. Kane and J. C. Browne. Using uncertainty in reputation methods to enforce cooperation in ad-hoc networks. In *Proceedings of the 5th ACM workshop on Wireless security*, pages 105–113. ACM, 2006.

E. Kochkina, M. Liakata, and A. Zubiaga. All-in-one: Multi-task learning for rumour verification. *arXiv preprint arXiv:1806.03713*, 2018.

S. Kumar and N. Shah. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*, 2018.

S. Kumar, R. West, and J. Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602, 2016.

S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th International Conference on Data Mining*, pages 1103–1108. IEEE, 2013.

M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. A. Roberts, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–218, 2013.

Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han. A confidence-aware approach for truth discovery on long-tail data. *Proceedings of the VLDB Endowment*, 8 (4):425–436, 2014a.

Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1187–1198. ACM, 2014b.

Q. Li, Q. Zhang, and L. Si. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1173–1179, 2019.

X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava. Truth finding on the deep web: Is the problem solved? In *Proceedings of the VLDB Endowment*, volume 6, pages 97–108. VLDB Endowment, 2012.

Y. Li, J. Gao, C. Meng, Q. Li, L. Su, B. Zhao, W. Fan, and J. Han. A survey on truth discovery. *ACM Sigkdd Explorations Newsletter*, 17(2):1–16, 2016.

Y. Liu, K. Li, Y. Jin, Y. Zhang, and W. Qu. A novel reputation computation model based on subjective logic for mobile ad hoc networks. *Future Generation Computer Systems*, 27 (5):547–554, 2011.

Y. Liu, X. Jin, and H. Shen. Towards early identification of online rumors based on long short-term memory networks. *Information Processing & Management*, 56(4):1457–1467, 2019.

S. Lyu, W. Ouyang, H. Shen, and X. Cheng. Truth discovery by claim and source embedding. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2183–2186. ACM, 2017.

J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1751–1754, 2015.

J. J. Michaelson, Y. Shi, M. Gujral, H. Zheng, D. Malhotra, X. Jin, M. Jian, G. Liu, D. Greer, A. Bhandari, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*, 151(7):1431–1442, 2012.

T. Mukherjee, B. Parajuli, P. Kumar, and E. Pasiliao. Truthcore: Non-parametric estimation of truth from a collection of authoritative sources. In *Big Data (Big Data), 2016 IEEE International Conference on*, pages 976–983. IEEE, 2016.

S. Ng, E. A. Collisson, A. Sokolov, T. Goldstein, A. Gonzalez-Perez, N. Lopez-Bigas, C. Benz, D. Haussler, and J. M. Stuart. Paradigm-shift predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics*, 28(18):i640–i646, 2012.

Q. V. H. Nguyen, C. T. Duong, T. T. Nguyen, M. Weidlich, K. Aberer, H. Yin, and X. Zhou. Argument discovery via crowdsourcing. *The VLDB Journal*, 26(4):511–535, 2017.

V. Oleshchuk and V. Zadorozhny. Trust-aware query processing in data intensive sensor networks. In *Sensor Technologies and Applications, 2007. SensorComm 2007. International Conference on*, pages 176–180. IEEE, 2007.

R. W. Ouyang, L. M. Kaplan, A. Toniolo, M. Srivastava, and T. J. Norman. Parallel and streaming truth discovery in large-scale quantitative crowdsourcing. *IEEE Transactions on Parallel and Distributed Systems*, 27(10):2984–2997, 2016.

J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 877–885. Association for Computational Linguistics, 2010.

J. Pasternack and D. Roth. Latent credibility analysis. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1009–1020. ACM, 2013.

E. O. Paull, D. E. Carlin, M. Niepel, P. K. Sorger, D. Haussler, and J. M. Stuart. Discovering causal pathways linking genomic events to transcriptional states using tied diffusion through interacting events (tiedie). *Bioinformatics*, 29(21):2757–2764, 2013.

K. Pelechrinis, V. Zadorozhny, V. Kounev, V. Oleshchuk, M. Anwar, and Y. Lin. Automatic evaluation of information provider reliability and expertise. *World Wide Web*, 18(1):33–72, 2015.

V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104*, 2017.

R. Pochampally, A. Das Sarma, X. L. Dong, A. Meliou, and D. Srivastava. Fusing data with correlations. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 433–444. ACM, 2014.

C. Polymorphisms. Distinguishing cancer-associated missense mutations from. *Cancer Res*, 67(2):465–473, 2007.

M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*, 2017.

T. J. Pugh, O. Morozova, E. F. Attiyeh, S. Asgharzadeh, J. S. Wei, D. Auclair, S. L. Carter, K. Cibulskis, M. Hanna, A. Kiezun, et al. The genetic landscape of high-risk neuroblastoma. *Nature genetics*, 45(3):279, 2013.

V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1589–1599. Association for Computational Linguistics, 2011.

K. Quinn, D. Lewis, D. O'Sullivan, and V. P. Wade. An analysis of accuracy experiments carried out over of a multi-faceted model of trust. *International Journal of Information Security*, 8(2):103–119, 2009.

J. Reimand and G. D. Bader. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Molecular systems biology*, 9(1), 2013.

T. Rekatsinas, M. Joglekar, H. Garcia-Molina, A. Parameswaran, and C. Ré. Slimfast: Guaranteed results for data fusion and source reliability. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1399–1414. ACM, 2017.

V. L. Rubin and T. Lukoianova. Truth and deception at the rhetorical structure level. *Journal of the Association for Information Science and Technology*, 66(5):905–917, 2015.

N. Ruchansky, S. Seo, and Y. Liu. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 797–806, 2017.

V. Shcherbakov, J.-F. Gayet, O. Jourdan, A. Minikin, J. Ström, and A. Petzold. Assessment of cirrus cloud optical and microphysical data reliability by applying statistical procedures. *Journal of Atmospheric and Oceanic Technology*, 22(4):409–420, 2005.

B. Shi and T. Weninger. Fact checking in heterogeneous information networks. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 101–102. International World Wide Web Conferences Steering Committee, 2016.

Y. Shpungin. Networks with unreliable nodes and edges: Monte carlo lifetime estimation. *International Journal of Applied Mathematics and Computer Sciences*, 4(1):168–173, 2007.

K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36, 2017a.

K. Shu, S. Wang, and H. Liu. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*, 2017b.

K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*, 2018.

T. Sjöblom, S. Jones, L. D. Wood, D. W. Parsons, J. Lin, T. D. Barber, D. Mandelker, R. J. Leary, J. Ptak, N. Silliman, et al. The consensus coding sequences of human breast and colorectal cancers. *science*, 314(5797):268–274, 2006.

Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes. The cosmic cancer gene census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer*, 18(11):696–705, 2018.

C. Song, C. Yang, H. Chen, C. Tu, Z. Liu, and M. Sun. Ced: Credible early detection of social media rumors. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*, 2017.

H. Tan, J. Bao, and X. Zhou. A novel missense-mutation-related feature extraction scheme for ‘driver’mutation identification. *Bioinformatics*, 28(22):2948–2955, 2012.

C. J. Tokheim, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, and R. Karchin. Evaluating the evaluation of cancer driver genes. *Proceedings of the National Academy of Sciences*, 113(50):14330–14335, 2016.

L. A. Urry, M. L. Cain, S. A. Wasserman, P. V. Minorsky, and J. B. Reece. *Campbell biology*. Pearson Education, Incorporated, 2017.

F. Vandin, E. Upfal, and B. J. Raphael. De novo discovery of mutated driver pathways in cancer. *Genome research*, 22(2):375–385, 2012.

D. A. Waguih and L. Berti-Equille. Truth discovery algorithms: An experimental evaluation. *arXiv preprint arXiv:1409.6428*, 2014.

W. Y. Wang. ”liar, liar pants on fire”: A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

X. Wang, Q. Z. Sheng, X. S. Fang, L. Yao, X. Xu, and X. Li. An integrated bayesian approach for effective multi-truth discovery. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 493–502. ACM, 2015.

X. Wang, Q. Z. Sheng, L. Yao, X. Li, X. S. Fang, X. Xu, and B. Benatallah. Truth discovery via exploiting implications from multi-source data. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 861–870. ACM, 2016.

D. Wodarz, A. C. Newell, and N. L. Komarova. Passenger mutations can accelerate tumour suppressor gene inactivation in cancer evolution. *Journal of The Royal Society Interface*, 15(143):20170967, 2018.

Y. Wu, P. K. Agarwal, C. Li, J. Yang, and C. Yu. Toward computational fact-checking. *Proceedings of the VLDB Endowment*, 7(7):589–600, 2014.

Y. Yang, Q. Bai, and Q. Liu. On the discovery of continuous truth: a semi-supervised approach with partial ground truths. In *International conference on web information systems engineering*, pages 424–438. Springer, 2018.

L. Yin, J. Han, W. Zhang, and Y. Yu. Aggregating crowd wisdoms with label-aware autoencoders. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1325–1331. AAAI Press, 2017.

X. Yin and W. Tan. Semi-supervised truth discovery. In *Proceedings of the 20th international conference on World wide web*, pages 217–226. ACM, 2011.

X. Yin, J. Han, and S. Y. Philip. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6): 796–808, 2008.

A. Youn and R. Simon. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics*, 27(2):175–181, 2011.

V. Zadorozhny and J. Grant. A systematic approach to reliability assessment in integrated databases. *Journal of Intelligent Information Systems*, 46(3):409–424, 2016.

D. Zhang and V. I. Zadorozhny. Fake news detection based on subjective opinions. In *European Conference on Advances in Databases and Information Systems*, pages 108–121. Springer, 2020.

B. Zhao and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. *Proc. of QDB*, 2012.

B. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. *Proceedings of the VLDB Endowment*, 5(6):550–561, 2012.

Y. Zheng, G. Li, and R. Cheng. Docs: a domain-aware crowdsourcing system using knowledge bases. *Proceedings of the VLDB Endowment*, 10(4):361–372, 2016.

A. Zubiaga, E. Kochkina, M. Liakata, R. Procter, M. Lukasik, K. Bontcheva, T. Cohn, and I. Augenstein. Discourse-aware rumour stance classification in social media using sequential classifiers. *Information Processing & Management*, 54(2):273–290, 2018.