Reputation Interpretation

a white paper from Rebooting the Web of Trust IX

by Arthur Brock, Kaliya Hamlin, Grace (Rebecca) Rachmany, and Jakub Lanc

ABSTRACT

This paper explores how to take a "reputation" trust graph with multiple characteristics and create actionable output. It is not focused on the automation of the process, which could be applied by human processing of information or through a programmed software implementations.

We are looking to categorize the types of reputation that people might have into general buckets, such as:

- Knowledge
- Emotional Intelligence
- Generalized Skills
- Reviews from others

The output that we would produce is specific to each type of community or decision. For example:

- Yes/no
- Graph or diagram
- Report
- Derivative reputation assessment. For example, in moving from one sport to another, some skills might be transferable and some not.
- Complex adjustment to the terms of engagement. For example, if you are offering a job to someone, the output might be a change in their job description.

The process may also be applicable to business entities and organizations, not just persons.

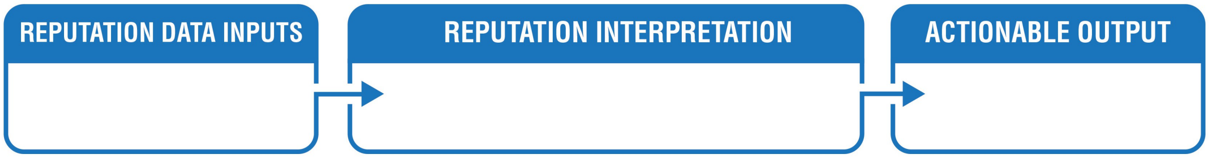
DCO

Protocol Labs

Sponsors for the Rebooting the Web of Trust IX Design Workshop

OVERVIEW

The fundamental assumption of this paper is that data collection is not the problem in creating a "reputation" for individuals, entities, or groups but that the issue comes when interpreting the data. Reputation inputs are constantly collected and understood differently in different contexts. We do not address the question of "what is reputation?" nor do we address "how to create interoperable reputation standards". Instead, the focus is on how to interpret reputation data.



In the above diagram, we visualize the process that we discuss in this paper. Reputation Interpretation, the box in the center, is the core of this paper. While this paper touches on Inputs and Outputs from the interpretation process, the focus of the paper is on the process an organization would undertake to Interpret the data in a way that makes sense to that organization. In that sense, the context is decisive.

Following is the process the authors identified as the Reputation Interpretation Process. The following description also serves as a mini-glossary to the terminology used in this document. Some of the terminology may be a placeholder as these concepts have not been well-defined previous to the paper. Once running, we imagine the process itself happening in the following chronological sequence.

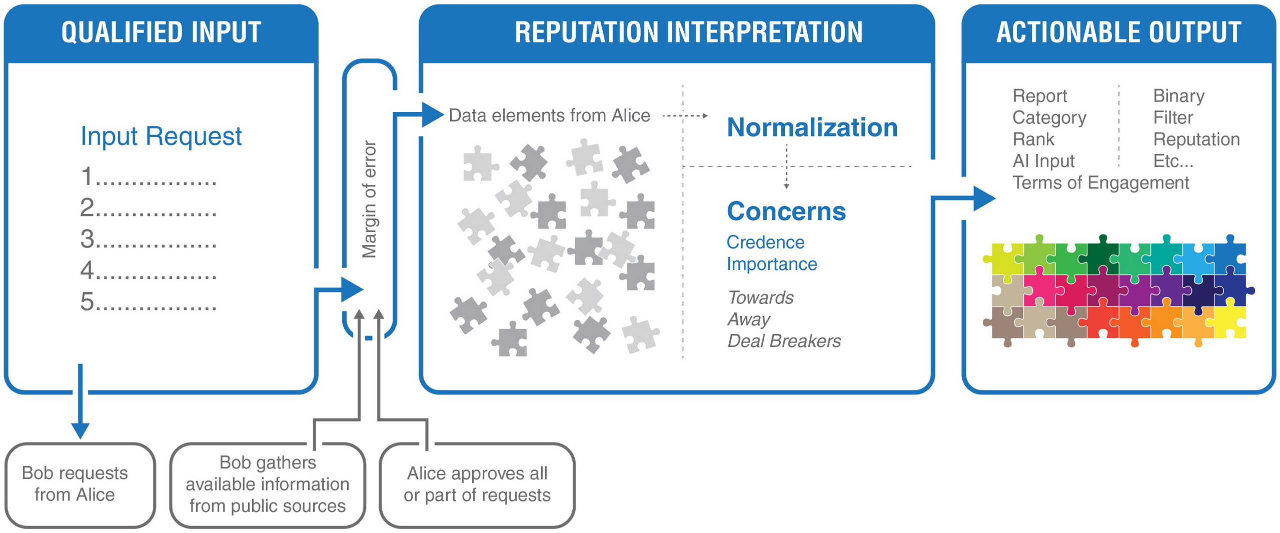


Illustration case: Throughout this discussion, we will use the case of interpreting data for mental health professionals, such as therapists or psychologists.

- **Input Qualification:** The Reputation Interpretation system user (Bob) defines the input desired for the specific context and output and sends a request to Alice for the information required. The data may come in different formats, discussed briefly below. For example, someone might ask a mental health professional for educational certifications, Myers-Briggs or other personality trait results, testimonials from clients, peer assessments, test results, etc.
- **Margin of Error:** Alice approves a certain subset of the information Bob requested. Bob has defined a "margin of error" based on less than 100% of the data. The Margin of Error represents how close to accurate the end result is. For example, if the health professional didn't send any testimonials, they may simply not have them or they may be hiding them, so the score needs to take this into account.
- **Normalization of Data Elements:** Data, even data of similar types (such as ratings or reviews) comes in different formats, so some form of normalization processing needs to be performed to put the data in a format that is understandable and processable by the system. Each piece of data might inform a different element of the output. For example, if the output is a recommendation of the types of patients Alice would serve the best, specific words identified in the reviews and assessments would correspond to the section describing Alice's specialty, for example, teenagers or recovering drug addicts.
- **Concerns:** Once the data is normalized, the specific concerns are applied to the data. The judgements are individual for each community or system. For example, Bob may think study of Jungian psychology is an advantage or a disadvantage for the health clinic.
    - Weights signify the importance of each element of data. For example, peer reviews are 30% of Alice's score.
    - Credence refers to how reliable the data is. For example, on a Myers-Brigg test result, Bob would consider the credence of the issuer and whether the test might have been gameable.
    - Towards, Away, and Dealbreakers. When considering the concerns, there are things that have positive (Towards) value, such as Bob wants people with proven experience. Other characteristics are negative (Away) values, such as Bob might think that if Alice has a certification as a Reiki healer, mental health patients in his facility might think it's less reliable. Dealbreakers are dealbreakers, for example, if Alice's drug tests show she is a heroin addict, Bob would not allow her to treat any patients.
- **Processing the Data:** Once the data is normalized, and the concerns are accounted for, the data can be processed to provide the final output.
- **Output:** The output of the Reputation Interpretation process is contextual and actionable. Output could be a number or yes/no answer, but in many systems it might be more complex. For example, Bob might offer Alice a job contract with customized terms of engagement, depending on the assessment of the data.

Designing the system would occur in the following order (which is reflected in the organization of this paper):

1. Define the output
2. Identify the inputs
3. Define normalization of the data
4. Define the concerns
5. Map the associations and calculations of the elements to produce the output

The remainder of the paper goes into more detail on each one of these steps.

DEFINE THE OUTPUT

Many discussions of reputation focus on numerical measures of reputation. Generally speaking, in the digital world, reputation is expressed as a kind of number, such as 5-star ratings, reputation points, credit rating, rankings, or accumulated reputation in the form of tokens. Within the digital identity community, credentials or claim are another common form of reputation. Alice may make a claim that she worked for a particular employer or holds a degree, and a yes/no verification of that claim by the institution in question indicates it is valid.

While these numbers and certificates are a convenient and accepted measures, and there has been a bias to think of reputation in these terms, numbers aren't necessarily actionable. When Bob chooses a therapist (for himself or his clinic), just knowing that Alice has 5-star ratings isn't enough. Perhaps Bob would want to read the experiences of other patients in long form as well, or perhaps Bob values professional peer opinions.

Likewise, there's been a bias towards thinking of "actionable" as making a binary decision. For example, Bob may be making a simple choice: purchase this service. However, Bob may be facing a complex choice, such as the terms of employment to offer Alice at his clinic, the workload and types of patients to assign to Alice, etc.

Following is a list of the types of outcomes discussed in the group. It is a starting list for the purposes of this paper and to allows Bob to consider varying types of outcomes. More suggestions for this list are welcome.

Types of actionable outcomes:

- Binary decisions (hire/no hire)
- Categorization (placing Alice in a category, such as a professional specialty)
- Matching recommendations (providing patients with a list of recommended therapists based on their needs)
- Terms of engagement or update to terms of engagement (Alice attended professional training and received a number of assessments, and it is time to review and Bob would like to update her work conditions)
- Choose from options (schedule an interview or session, purchase a subscription, buy a book or online course Alice has published)
- Filter (provide a list of the top candidates for a job offer)
- Assessment (report or other long-form assessment, such as in a personality test)
- Context-appropriate reputation (translate someone's ranking or reputation points in one context to another context)
- Utilization for machine learning (assessing effectiveness of different therapeutic methods on different profiles of patients)
- Ranking
- Disqualification

The different types of output are appropriate for different contexts and intent. In defining the output, other areas to consider are:

- **Intent:** What is done with the output, how long it is retained, whether this is a one-time decision or part of a larger decision.
- **Relevant Actions:** The actions that will be taken based on the output.
- **Decision Criteria:** The important factors for making decisions. This varies on the context and judgment of the community or individual making a decision. For example, Bob may consider professional assessment as extremely important.
- **Output Requirements:** The format in which the output will be presented. For example, Bob may get one report for his decision-making regarding the terms of engagement for Alice's work contract. A patient at Bob's clinic may receive a different presentation of the output. Some of the information about Alice may never be exposed to Bob. For example, Alice's gender, age, and race would not be exposed, although that information is in the system, in order to avoid biases.
- **Individualization of Outputs:** In some contexts, Bob may want to provide individualization of outputs. For example, IQ tests are biased towards certain racial and educational backgrounds. Bob may want to create individualized reports taking into account such biases, or individualized reports based on different educational backgrounds. For someone with a strong academic background, ongoing certifications might be less important than someone who didn't do as much initial study.

Illustration case: Bob's output will be a contract for employing Alice, as well as matching recommended clients, and creating a short overview that is shown on the clinic's website and that patients review to help them choose their practitioner.

IDENTIFY INPUTS

Once the output is well-defined, it is possible to determine the inputs that can be used to reach that output. Starting with the end in mind immediately simplifies the seemingly impossible tasks of wading through the tremendous amount of information available.

This paper makes no attempt to define or categorize the types of inputs available, nor does it make any suggestions about how that input should be provided.

Following are the considerations for Bob when identifying and determining inputs.

- **Ideal:** The ideal outcome would be perfect in every way. Bob would have all of Alice's academic record, information on every outcome for every patient Alice had ever served, peer reviews of her work, patient reviews, etc. All of that "perfect" information would be 100% reliable and not gameable.
- **Available:** In the real world, not all information is available. For example, Bob ideally would like to know the symptoms or complaints of each client who went to Alice, how long they were treated for, and what

the outcome was. For example, Alice might have taken three hours to eliminate someone's phobia, or she might have treated them for three years, but they still have symptoms. While ideally, Bob would have access to this information about every patient Alice saw, in fact, none of it is generally available. Even in a future world where the patient could release that (anonymized data), some of Alice's clients would release the information and some would not, so it would be incomplete.

- **Cost:** Some data costs money, whether in terms of the actual purchase of the data, or in the form of difficulty of processing that data.

Following these considerations, Bob would end up with a list of the data and data sources that Bob will request from Alice.

MARGIN OF ERROR

Bob now makes a request from Alice for the data, and Alice provides all or a subset of the data. Looking at the subset of data, Bob may assess the "Margin of Error". Bob may or may not know whether Alice simply doesn't have that data (for example, she never had a Myers-Brigg assessment performed), or whether she intentionally withheld data (for example, omitting an employer from her resume or LinkedIn profile). Based on the completeness or incompleteness of the data, Bob determines the "reliability" or "margin of error" for the final score.

Some data may be more critical than others in determining "margin of error." For example, if Alice is missing test results from academic institutions, Bob might consider that a minor factor for margin of error, but if she is missing actual certification from the institution, that might be a major omission that Bob considers giving a high "margin of error" because her claim to have graduated now seems dubious, putting into doubt all the rest of her self-attestations.

The authors of the paper were divided as to whether the Margin of Error would be calculated as an "overall" score on the report, or whether it would be applied individually to each item of data. The sequence of the calculation is an implementation issue. This paper points out the issue of completeness of data as one to take into consideration, and it is up to the Bobs of the world to determine where they want to implement it.

NORMALIZATION

Once data is collected, the data needs to be normalized to make it understandable and able to be processed. Following are considerations for normalization of data.

- Data may be provided in difficult-to-understand or variable formats. For example, ratings might use a numerical 1-5 scale or a 1-10 scale (odd scales have a middle option and even scales do not), a thumbsup-thumbsdown rating, or a multifaceted rating (professionalism, effectiveness, friendliness, good-for-families, etc.) A system collecting professional assessments or reviews may need to normalize the data so that it's all provided on a similar scale.

- Data may be gameable. The system itself may have vulnerabilities: people may be able to put in false reviews, hack the system, be biased for certain behaviors, etc.
- User control: Some institutions will provide uncensored data. For example, a university transcript includes all grades for all courses, and Alice cannot edit out anything. In LinkedIn, when someone provides a recommendation, Alice can approve or decline, or request a change to the review. This is not "gaming" the system, just the fact that each system offers different levels of control by the user.
- Trustworthiness (reputation in the non-formal sense of the word) of the issuer: A grade from a trade school may be worth more or less to Bob based on the institution's reputation, how much Bob values applied learning versus academic learning, etc.
- Biases on the type of data: For example, review systems tend to be biased towards excellent and horrible reviews, because of human bias. If people have an okay or neutral response, they tend not to provide any review. By definition, human reviews tend towards the extreme.

The above are top-level concerns for unpacking the data. Furthermore, data may be multidimensional. For example, records from Alice's graduate studies might include grades, peer reviews, student reviews, number of publications, letters of recommendation, awards, etc.

Part of the normalization process is attributing the different characteristics to the appropriate part of the outcome. For example, if Alice worked for a particularly prestigious institution, that would affect the salary offered and that institution's name would be mentioned in the website blurb.

CONCERNS

Once data from different sources has been normalized, the data is presumably now n a format that allows similar types of data to be processed together in groups. For example, peer reviews would be normalized as positive, neutral, or negative. Text that will be scanned for keywords indicating specialties will be normalized into categories separating client reviews from peer reviews from employer recommendations.

In Bob's example, two types of concerns were identified:

- **Importance:** Bob will use weighting mechanisms to determine the importance of the elements in determining output. For example, patient reviews may be an important part of what potential clients see, taking up 50% of that report, but in providing the contract of employment, Bob may give them no weight, and just have a threshold, for example, hire Alice if she has more than 80% positive reviews.
- **Credence:** Each data source (or aggregated data element) has a level of credence for that particular value. The credence is an assessment of how likely it is that the data is believable. The believability of the data can be influenced by anything from the issuer through the hackability of the connection when the data was provided. Each element can be weighted according to its credence.

PROCESSING

Bob's on his own here. Maybe he'll attend RWoT10 and add this part.

CONCLUSIONS

The discussion of how to measure reputation, what reputation is, and how to track it becomes an endless rabbithole because of the huge range of aspects of reputation and its measurements that could be discussed. By taking a practical approach, that is, discussing reputation in a specific context, we are able to understand that reputation isn't a thing unto itself, but a tool that people and organizations use in order to make decisions. By beginning with the end in mind, we are able to make sense of reputation, both from the perspective of understanding what data to gather about a person, and how to process it.

This paper provides a framework for how to think about reputation from the perspective of the actions and decisions that are made based on the reputation. For each context, different measures are brought in, normalized, and assessed in order to make the relevant decision.

Additional Credits

Lead Author: Grace (Rebecca) Rachmany

Authors: Arthur Brock, Kaliya Hamlin, and Jakub Lanc

Sample APA Citation:

Brock, A., Hamlin, K., Rachmany, G., and Lanc J. (2019). Reputation Interpretation. *Rebooting the Web of Trust IX*. Retrieved from https://github.com/WebOfTrustInfo/rwot9-prague/blob/master/final-documents/reputation-interpretation.pdf.

About Rebooting the Web of Trust

This paper was produced as part of the Rebooting the Web of Trust IX design workshop. On September 3rd to 6th, 2019, over 60 tech visionaries came together in Prague, The Czech Republic to talk about the future of decentralized trust on the internet with the goal of writing at least 5 white papers and specs. This is one of them.

RWOT Board of Directors: Christopher Allen, Joe Andrieu, Kim Hamilton Duffy

Members of the Organizing Committee: Dan Burnett, Dmitri Zagidulin

Sponsors: Digital Contract Design (Gold), Protocol Labs (Silver), Digital Bazaar (Ongoing Sustaining), Jolocom

Community Sponsors: Blockchain Commons, ConsenSys, Learning Machine, Legendary Requirements

Workshop Credits: Christopher Allen (Founder, Co-Producer), Joe Andrieu (Co-Producer and Facilitator), and Shannon Appelcline (Editor-in-chief).

Thanks to our other contributors and sponsors!

Thanks also to Paralelní Polis and the Institute for CryptoAnarchy in Prague.

What's Next?

The design workshop and this paper are just starting points for Rebooting the Web of Trust. If you have any comments, thoughts, or expansions on this paper, please post them to our GitHub issues page:

> https://github.com/WebOfTrustInfo/rwot9/issues

The tenth Rebooting the Web of Trust design workshop is scheduled for early 2020. If you'd like to be involved or would like to help sponsor the event, email:

> rwot-leadership@googlegroups.com