
From Personal Genomics to Twitter: Visualizing the Uncertainty of Evidence

Lauren Westendorf

Christina Pollalis

Clarissa Verish

Orit Shaer

Wellesley HCI Lab
Wellesley College
Wellesley, MA, USA
{lwestend, cpollali, cverish,
oshaer}@wellesley.edu

Panagiotis Takis Metaxas

Wellesley College
Wellesley, MA, USA
pmetaxas@wellesley.edu

Samantha Finn

Wellesley College
Wellesley, MA, USA
sfinn@wellesley.edu

Madeleine Ball

Open Humans Foundation
Brookline, MA, USA
madeleine@openhumans.org

Oded Nov

New York University
New York NY, USA
onov@nyu.edu

Abstract

Personal genomics offers a complex form of uncertainty in which a person's data are largely stable, but the interpretation and implications continue to evolve with the emergence of new research. Another domain, in which there is uncertainty about the supporting evidence and truthfulness of a claim, is social networks. We propose that a similar method can be used to communicate uncertainty in these contexts, and present a tool for visualizing social network claims that builds upon research in both contexts.

Author Keywords

Uncertainty; personal genomics; personal informatics.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

Background

The personal genomic context offers a form of uncertainty not addressed by existing taxonomies and applications. Although personal genomic data are largely stable during a person's lifetime, the interpretation and implications of such data change over time as advances in medical research uncover relationships between genes and health. In traditional forms of personal informatics, uncertainty often arises from the accuracy and context of data tracking. In

contrast, in genomic data 99.9999% accuracy is typical, and continues to improve [4]. However, the interpretation of the scientific evidence and its certainty, and the related implications for the user's health, frequently evolve as new medical research reveals new relationships between a person's genetic makeup and resulting effects.

To address uncertainty in this context, in previous work, we developed GenomiX—an interactive visual tool to support self-exploration of personal genomic data [5], which helps users understand the interpretations' level of certainty. GenomiX does not provide any new genetic interpretations, but instead provides new and easier ways for users to explore the data, drawing upon the publicly available interpretation provided by the Personal Genome Project (PGP) [3]. In GenomiX, gene variants are represented as bubbles plotted by potential health effect (low, medium, high) and the certainty of the scientific evidence (well-established, likely, uncertain) that links a gene variant with a trait or condition. The color and size of each bubble encode effect (pathogenic, benign, protective) and risk. These designations are drawn from the PGP database and are updated as new evidence is discovered.

Encouraged by the positive acceptance of the tool by users who explored their own data [5], we seek to apply our approach to help non-expert users in domains beyond personal genomics to better understand the uncertainty associated with the interpretation and credibility of information. One domain, in which there is often uncertainty about the credibility of information, is social networks where the quality of supporting evidence and truthfulness of a claim is not clear. Users encounter stories shared by

others in their network and are often unaware of the uncertainty regarding the trustworthiness of these stories. Following, we describe the application of GenomiX's interactive visualization approach to communicating the uncertainty regarding the credibility of claims spread on Twitter using Twitter Trails [6].

Twitter Trails

TwitterTrails is an investigative and exploratory tool to analyze the origin and spread of a "story" on Twitter (a claim, a meme or an event). Prompted by a search with relevant keywords using the Twitter API, it collects and analyzes up to 200K tweets automatically. While it does not answer directly the question of a story's validity, it provides information that a critically thinking person can use to examine how a Twitter audience reacts to the spreading of the story.

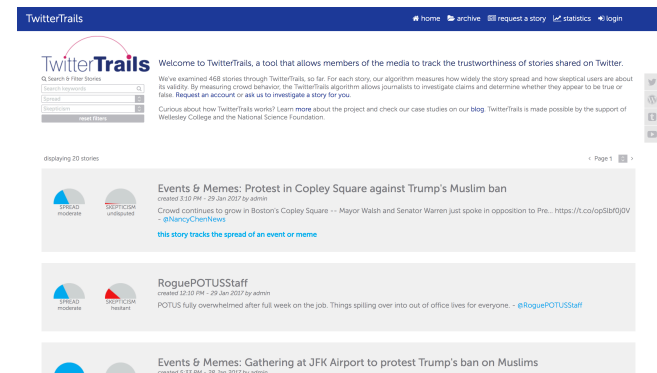


Figure 1. TwitterTrails homepage.

We used TwitterTrails to infer certainty and truthfulness of claims spread on Twitter using the following measures:

Spread measures simultaneously the propagation and number of the highest reaching tweets [2]. It does not reflect the few tweets with the most retweets, nor does it measure how many tweets were collected. Although these numbers are interesting and meaningful, the spread is meant to give an overall picture of the impact of a story: how visible it was as well as how many people were engaged in it.

Skepticism measures the prominence of doubt and mistrust in a story [2]. The first step to calculating the skepticism is to identify tweets in which the author expresses doubt in the validity of a claim, whether they are wondering if the claim is false or expressing that it is an outright lie. For now, TwitterTrails employs a simple algorithm, which works fairly well: identify tweets containing commonly used keywords to express doubt or disbelief (such as “hoax” or “fake”), which can be modified for individual stories and languages. Using this algorithm, data is separated into two subsets: those that express doubt and those that do not (implicitly expressing support). The skepticism of a story is defined as the ratio of the spread of doubting tweets to the spread of supporting tweets.

Partisan bias is computed from characteristics of the co-retweeted network [1] that is formed using a force-directed drawing algorithm using Gephi. This is done by measuring the size, density, and Euclidean distance between polarized groups of retweeters, while their political identity (liberal or conservative) is determined

by the frequency of politically-charged keywords appearing in the group members' profile description.

TwitterTrails' current database of about 500 investigated stories offers evidence that large enough crowds following a story are likely to react in ways that strongly correlates with the validity of a claim. True claims are likely to have relatively high spread with low skepticism, while false claims typically have low spread and high skepticism.

Claims that have both low spread and skepticism are difficult to categorize, and we have found almost no examples of claims that have both high spread and high skepticism.

Visualizing Twitter Claims

Using data from TwitterTrails, we can plot claims and events on Twitter in a similar manner to the gene variants in GenomiX (Figure 2), where the x- and y-axes represent the skepticism score and truth of the claim and the size of each bubble corresponds to the spread of the claim. The color encodes the partisan bias of the users discussing the claim, where claims discussed in a partisan echo chamber are more saturated and claims with more bipartisan discussion are gray or white. When a user selects a claim, detailed information about that claim (taken from the Overview in TwitterTrails) appears in the right sidebar, including a summary of the claim, the number of tweets and users discussing it, keywords, and the collection date.

Visualization: TwitterTrails

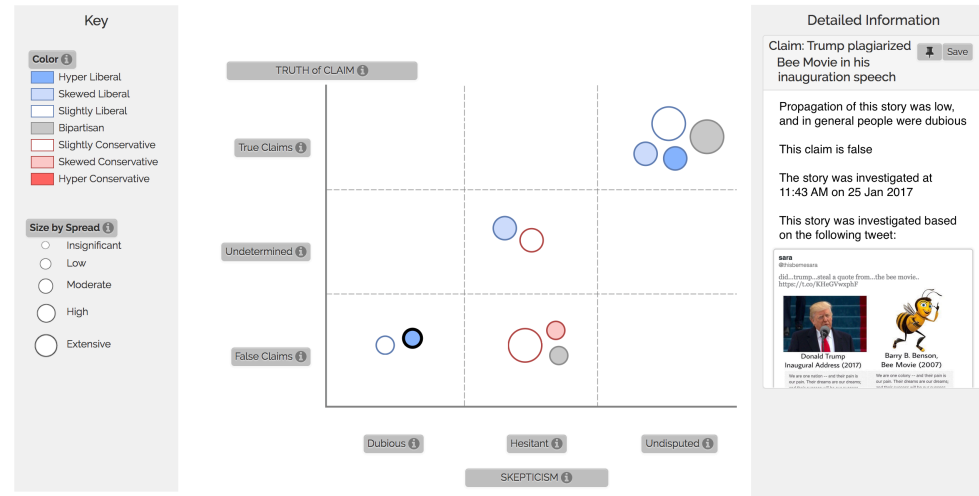


Figure 2: Visualizing TwitterTrails stories. Each bubble represents a twitter claim or event, plotted by truth of claim and skepticism. When a claim is selected, detailed information about that claim is displayed on the right.

Conclusion and Future Work

Visualizing data which reduce uncertainty is a challenge across many domains. In this paper, we used personal genomics and social media claims as unique and compelling contexts for communicating uncertainty. We also proposed extending visualization techniques from personal genomics to map the spread and skepticism of twitter claims using data from TwitterTrails. Our previous work in the domain of personal genomics aimed to encourage exploration and long-term engagement for non-expert users through the development and evaluation of a novel visualization tool. We extend this goal and hope to empower consumers of the news and social media users to better understand uncertainty associated with the credibility of

stories. We also aim provide such users with visual tools which enable and encourage them to explore the propagation and supportive evidence of stories. Future work will include evaluation of the usability, usefulness and impact of our proposed tool. We believe that enabling users to map retweeted claims from their own account could lead to self-reflection about their role in misinformation propagation.

Acknowledgements

We thank our collaborators from Open Humans. This work is partially funded by NSF grant IIS-1422068.

References

1. Samantha Finn, Panagiotis Takis Metaxas and Eni Mustafaraj. "Measuring Perceived Political Polarization through Collective Intelligence presented". Proceedings of the Conference on Political Networks, 2014.
2. Samantha Finn, Panagiotis Takis Metaxas and Eni Mustafaraj. 2015. Spread and Skepticism: Metrics of Propagation on Twitter. in Proceedings of the ACM Web Science Conference, ACM, 39.
3. Personal Genome Project. 2016. Retrieved January 31, 2016 from <http://www.personalgenomes.org/>.
4. Brock A Peters, Bahram G Kermani, Andrew B Sparks, Oleg Alferov, Peter Hong, Andrei Alexeev, Yuan Jiang, Fredrik Dahl, Y Tom Tang and Juergen Haas. 2012. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature*, 487 (7406). 190-195.
5. Orit Shaer, Oded Nov, Johanna Okerlund, Martina Balestra, Elizabeth Stowell, Lauren Westendorf, Christina Pollalis, Jasmine Davis, Liliana Westort, and Madeleine Ball (2016). "GenomiX: A Novel Interaction Tool for Self-Exploration of Personal Genomic Data." Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 661-672.
6. Twitter Trails. 2016. Retrieved January 31, 2016 from <http://twittertrails.com/>.