Common Sense Reasoning about Source Credibility

Peiyao Hu (phu10@stevens.edu)
Mark K. Ho (mho4@stevens.edu)
Department of Computer Science, Stevens Institute of Technology
Hoboken, NJ, 07030, United States

Abstract

We often rely on others' testimony to learn about new topics, such as health benefits of a novel food. However, the sources are not always knowledgeable, helpful, or unbiased, necessitating an assessment of their credibility. Here, we present a Bayesian model of source credibility, where a listener reasons about the expertise and intentions of a source. We consider a scenario where both parties have noisy access to the ground truth of familiar topics, which serves as a basis for rational inference about a source's credibility on novel topics. This approach provides a computational framework for understanding how people respond to information in domains like science communication and media consumption.

Keywords: Social Learning; Social Epistemology; Theory of Mind; Probabilistic Programming; Bayesian Modeling

Introduction

We often rely on testimony from others when learning about new topics. Recommendations from social media, for example, can inform our choice on whether to include a new food into our diet. Given a constant influx of information from various sources, we face the problem of deciding which sources are credible and what messages to believe. At a minimum, we need to discern whether a source is knowledgeable, whether it communicates its genuine beliefs, or whether it has concealed motives. To effectively judge these factors, examining a source's views on familiar topics can be useful. This can serve as a basis for listeners to infer the credibility of the sources, which can be generalized to evaluate their testimony on novel topics.

Reasoning About Credibility

We present a Bayesian account of reasoning about source credibility, where the listener simultaneously infers the expertise and intention of a source while trying to figure out the truth. Previous research has shown that adults consider the expertise and potential biases of sources when estimating the value of used cars (Birnbaum & Stegner, 1979) while young children readily infer whether the informant is knowledgeable and helpful in epistemic trust tasks (Landrum, Eaves, & Shafto, 2015). Therefore, we consider how the credibility of a source can be derived from inference about three attributes: knowledgeability (being an expert), helpfulness (being informative), and bias (having persuasive goals).

Model Framework We model a simplified setting involving one source and one listener, where there is a familiar topic (e.g. the healthiness of broccoli) and a related, novel topic (e.g., the healthiness of avocado). The source has information about both topics, while the listener only has prior experience with the familiar topic. The source offers suggestions on these topics by proposing a discrete rating ranging from 0 to r_{max}, without any background of the source itself being available. The familiar topic allows the listener to reason about the source's knowledgeability, helpfulness and bias based on a kernel of correlated information (Figure 1).

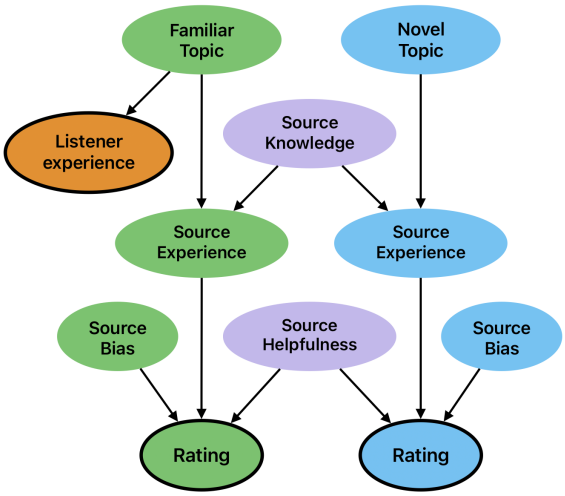

Figure 1: Bayesian network graph of reasoning about source credibility. Source knowledge and helpfulness are shared across topics while the bias is topic specific. The information available to the listener is marked by bold contours.

The truth value or goodness of a topic p ranges from 0 to 1. For the familiar topic, both parties have noisy access to the ground truth through probabilistic sampling, which yields either positive or negative outcome. The number of samples a source can draw is determined by its knowledgeability. For example, drawing N samples would lead to N_+ positive samples according to $\text{Binomial}(N, p)$. We refer to the accumulated outcomes as one's experience.

Recursive reasoning Our model builds on the rational speech act framework (Goodman & Frank, 2016), in which Bayesian agents reason about one another's decisions and inferences recursively: A savvy listener L with experience e interprets the intended meaning and goals that would have led a source S to produce a particular rating r

$$P_L(p|r,e) \propto P_S(r|p)P_L(p|e).$$

Meanwhile the source reasons about a naive listener L_0 who simply accepts the rating at face value

$$P_{L_0}(p|r) \propto \delta(r/r_{max}, p)P(p),$$

where δ returns 1 if the normalized rating equals to the true value of the topic and 0 otherwise.

The rating proposed by the source is generated by maximizing a combined utility, consisting of a persuasive utility U_p that bias the listener toward perceiving the truth as either positive or negative, and an informational utility U_i that seeks to convey the source's own beliefs P_S as accurately as possible

$$U_p = \mathrm{E}_L[p] - 0.5,$$
$$U_i = \mathrm{S}(P_S, P_L),$$
$$U = h * U_i + b * U_p,$$

where the discrete weighting factors are $h \in \{0,1\}$ and $b \in \{-1,0,1\}$. The S measures the similarity between beliefs

$$\mathrm{S}(P_1, P_2) = 1 - \mathrm{D}(P_1, P_2),$$

where D quantify the Wasserstein distance of two distributions, where the minimum distance is 0 for identical ones. The distributions in our case are defined within $[0,1]$, thus the largest distance is 1.

Quantify Credibility We formalize a multi-dimensional notion of credibility in an agent-relative manner, such that a fully credible source is one that would lead its listener to draw conclusions identical to those of the naive listener. The latter can be thought of as instantiating a form of trust which involves a suspension of the deliberative process (Nguyen, 2022).

Specifically, we quantify source credibility through the similarity $\mathrm{S}(P,Q)$ between the posterior P of the savvy listener and posterior Q of a native listener after hearing the source's testimony about a novel topic. The credibility C is defined as the averaged similarity in response to all possible ratings, minus the baseline credibility C_0 of an unknown source appears to the savvy listener

$$C = \frac{1}{r_{max}} \sum_{r}^{r_{max}} S(P_r, Q_r) - C_0.$$

Simulation Results We modeled two situations where the listener's belief on the familiar topic is either neutral or polarized (Figure 2). We manipulated the certainty in those beliefs by varying the number and outcome of samples drawn by the listener. This allows us to assess the impact of the listener experience on the perception of the source credibility.

The source is either not knowledgeable or an expert (can draw 1 or 10 samples, respectively). For the opinionated listener who happens to consistently draw positive samples, credibility increases if the source suggests the highest rating for the familiar topic, while eventually decreases for other ratings (Figure 2, top). For the moderate listener with mixed experiences, the credibility is lower than baseline for extreme ratings but higher for intermediate ratings (Figure 2, bottom).

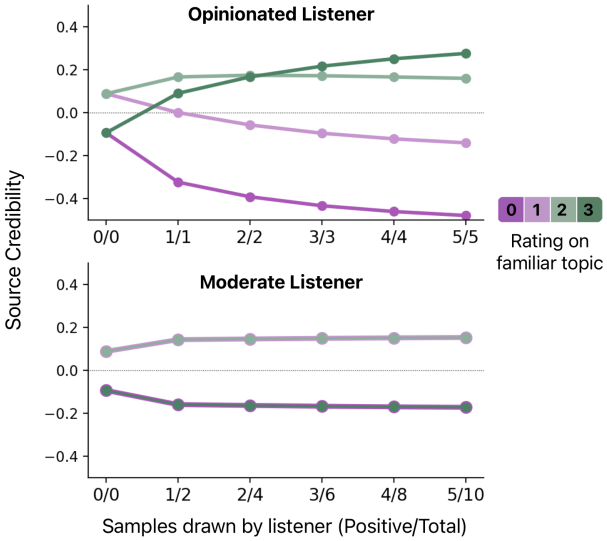

Figure 2: Source credibility assessed by listeners with varying experience.

Discussion

In this study, the listener models the source with domain general helpfulness and knowledgeability. Generalizing different attributes of the source across topics or social groups can facilitate social learning but might sometimes lead to undesirable consequences. We plan to apply this general computational framework to study how people conduct common sense reasoning about source credibility in contexts such as science communication and media consumption, as well as extend it to study more complex belief networks at both individual and collective levels (Vlasceanu, Dyckovsky, & Coman, 2024).

References

Birnbaum, M. H., & Stegner, S. E. (1979). Source credibility in social judgment: Bias, expertise, and the judge's point of view. *Journal of Personality and Social Psychology*, *37*(1), 48.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Landrum, A. R., Eaves, B. S., & Shafto, P. (2015). Learning to trust and trusting to learn: A theoretical framework. *Trends in Cognitive Sciences*, *19*(3), 109–111.

Nguyen, C. T. (2022). Trust as an unquestioning attitude. In *Oxford Studies in Epistemology Volume 7*. Oxford University Press.

Vlasceanu, M., Dyckovsky, A. M., & Coman, A. (2024). A network approach to investigate the dynamics of individual and collective beliefs: Advances and applications of the bending model. *Perspectives on Psychological Science*, *19*(2), 444–453.