Evaluation without Ground Truth in Social Media Research

Reza Zafarani and Huan Liu, *Arizona State University*

With the rise of social media, user-generated content is available at an unprecedented scale. These massive collections of user-generated content can help researchers better understand the online behavior of billions of individuals. This data also enables novel scientific research in social sciences, anthropology, psychology, and economics at scale.

Scientific research demands reproducible and independently verifiable findings. In social media research, scientific findings can be in the form of behavioral patterns, as in Individuals commented on this Facebook post due to its quality. To validate these patterns, researchers can survey the individuals exhibiting a pattern to verify if it truly captured their intentions. In data mining terms, such validation is known as evaluation with ground truth. [19]. However, social media users are scattered around the world. Without face-to-face access to individuals on social media, is it even possible to perform an evaluation in social media research? That is, how can researchers verify the user behavioral patterns found are indeed the true patterns of these individuals?

This problem is known in the data mining and machine learning literature as lack of ground truth, or lack of a gold standard that can help validate a scientific hypothesis. With so many opportunities for novel methods and findings on social media and limited ground truth, there is a pressing need for evaluation without ground truth in social media research. To address this problem systematically, researchers must consider the types of questions asked in social media research.

In social media research, researchers are often interested in when and where certain user activity is likely to take place and, if possible, why it takes place; for instance, a researcher might be interested in when individuals are more likely to tweet or search for restaurant reviews, at, say, a restaurant or home. Researchers also seek answers to why questions on social media: Why are users abandoning site A for site B?, and Why do users like Twitter despite the limit on the number of characters in a tweet? It is through such questions researchers look for inklings of causality in social media. They are also interested in how an algorithm or incentive mechanism will work prior to its public release on social media. Companies often face similar questions when evaluating their

methods on social media, from assessing the effectiveness of a new friend-recommendation algorithm to predicting the success of a new rewards program prior to release. Relevant recommendations and appropriate incentives can help increase user engagement and retention rates and, ultimately, sustain or increase profit.

Consider predicting when or where a particular user activity is going to happen. Unlike Apollo, the Olympian deity with knowledge of the future, humans find it a challenge to design methods able to predict the future. To answer, researchers design data-mining techniques that predict the most likely place or time an activity will happen in social media. The challenges introduced by humans' lack of knowledge about the future are further compounded by yearning to understand why things happen on social media. Without surveying users on social media, the gap between personal understanding and reality cannot be gauged.

Here, we discuss three types of evaluation for social media research that stand up to scientific scrutiny:

- **Spatiotemporal**. In response to validating our discoveries on when or where things are going to happen, or spatiotemporal predictions, in social media;
- **Causality**. Evaluating our hypotheses on causality, or why things are happening in social media; and
- **Outcome.** Outcome-evaluation techniques assess how well a computational method (such as an algorithm, application, or incentive mechanism) predicts an outcome or finds patterns; outcome evaluation also helps determine how to improve the computational method to perform better.

Practitioners and researchers alike in various fields, including statistics, computer science, sociology, psychology, epidemiology, and ethology, have developed a range of methods social media researchers can borrow and tweak in their search for reproducible evaluation methods for social media research. We present the most promising such methods, illustrating how they can be applied to social media research when ground truth is unavailable for evaluation.

1 Spatio-Temporal Evaluation

Consider designing a method that predicts the most likely time users will check their email messages or the restaurant they will most likely choose for dinner using their checkins, or personally reported locations, in social media. As the time or place predicted by the method occurs in the future, evaluation is a challenge. One straightforward heuristic that helps evaluate such spatiotemporal predictions is that individual behavior is periodic; for example, if an individual has checked email at a certain time for the past two days, it is then likely the same pattern will be observed in the same individual today (see Figure 1). The periodicity of human behavior simplifies evaluation for spatiotemporal predictions in social media.

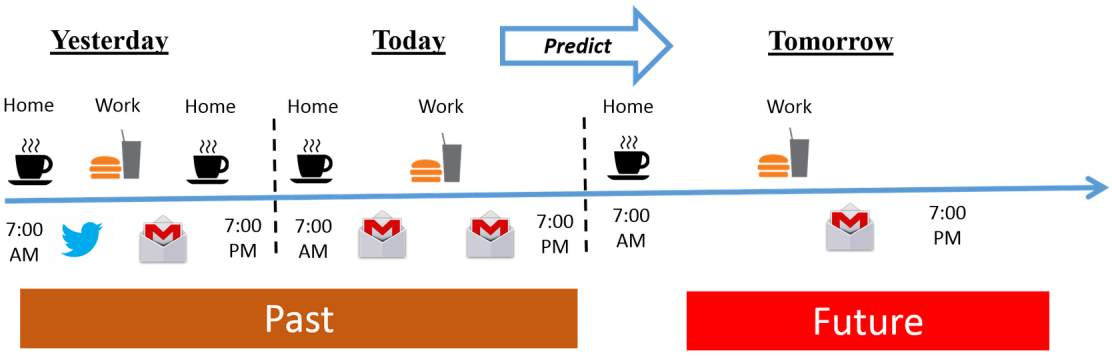

Figure 1: *Spatiotemporal evaluation without ground truth exploits the periodicity of human behavior.*

Spatiotemporal predictions can be performed using a machine-learning method. Researchers collect historical data about an individual for a period of time. Following the common practice in machine learning, the data collected in this period is partitioned into two consecutive periods of training and testing; for example, for 10 days of data, the first seven days is used for training and the next three days for testing. The machine-learning method is provided with the training data, and its outcome is evaluated on the testing data. If the method performs well on the testing data, the researcher can safely assume that, due to the periodical nature of human behavior, it will perform well in the near future. Note we assume human behavioral patterns are consistent over time, which is often not true for long periods. The researcher must therefore train the machine learning method repeatedly over time to ensure the technique consistently performs well.

What if a spatiotemporal prediction is for some behavior for which the researcher cannot guarantee recurrence, or periodicity, over time? One approach is crowdsourcing, using the wisdom of the crowd. In it, researchers ask opinions of multiple experts on a real-world problem (such as When will some phenomenon happen?) and choose the solution the majority agree on. Online labor markets, including Amazon's Mechanical Turk (http://www.mturk.com/), can provide a low-barrier entry to hiring individuals for performing crowdsourced evaluations on social media research. [3, 14].

Ensemble methods [5] are a counterpart of crowdsourcing in machine learning; for instance, consider predicting whether or not someone will tweet. Researchers can employ statistically independent methods to predict whether or not an individual will tweet. These methods can be combined by considering the prediction to be what the majority of methods predict will happen, or that the user will or will not tweet. Because methods are independent, their predictions should be independent of one another. If more independent methods agree on a solution, the solution is likely a valid solution to the problem [19, 5, 22].

2 Causality Evaluation

Consider how to determine the cause of a sudden increase in network traffic on a social media site. The researcher can speculate that some breaking news on the site is attracting users, resulting in a sudden increase in traffic. To determine the true cause of the traffic, one approach is to identify the counterfactual, or, in this case, what would have happened if the breaking news had not been released. That is, are there any other possible explanations (such as a malicious denial-of-service attack on the site) that could explain the heavy traffic? If no other explanation can be identified, then the researcher can safely say the breaking news is causing the traffic.

However, since determining the counterfactual requires knowing all other explanations, the researcher can resort to a controlled experiment if investigating the counterfactual is a challenge.

2.1 Controlled Experiments

In a controlled experiment, users are randomly assigned to two groups: control and treatment. A treatment is administered to the treatment group, while the control group receives no treatment. If the treatment results in a significant outcome in the treatment group, then the researcher can safely say the treatment is causing the outcome. In our network-traffic example, the researcher releases the breaking news as a treatment to a random set of users (treatment group). The researcher then compares the level of traffic in this set to that of another random set of users for whom the breaking news is not shown (control group). If the traffic is significantly heavier in the treatment group, the researcher can conclude the breaking news is causing the traffic.

Note in a controlled experiment, the counterfactual is approximated without considering all other possible explanations. However, when taking random populations of users, a researcher is taking samples of all other explanations and comparing them to the treatment group that is shown the breaking news. To improve the confidence of controlled experiments, Sir Ronald A. Fisher proposed randomized experiments in which the researcher takes many random treatment populations and observes if the heavy traffic is observed in more than $(1 - p)\%$ of the treatment groups; p denotes the confidence level, and a value of $p = 0.05$ is often considered in practice.

Controlled experiments are shown to be highly effective in social media research. As an example, La Fond et al.. [13] demonstrated how controlled experiments can be used to determine if influence is causing users to change behavior (such as by adding new hobbies due to a friend's influence). In these experiments, they generated control groups by randomizing user attributes (such as interests) over time. They assumed if influence exists, the influencer should become more similar to the influencee over time and this increase in similarity should be greater among influenced users than among randomly generated control groups.

Conducting controlled experiments is not always straightforward and can be financially and computationally expensive, and, at times, even impossible. So, a researcher must seek an alternative to find the counterfactual. A natural experiment is one such alternative.

2.2 Natural Experiments

Nature often provides researchers and ordinary people alike a randomized experiment with no extra effort; for instance, consider a city in which a researcher is analyzing whether an increase in number of police officers will reduce the rate of street crime. To perform controlled experiments, the researcher must randomly select some neighborhoods, observe the change in crime rate after officers are assigned to them, and compare this change to those of other neighborhoods where officers are not assigned. However, the researcher cannot afford to perform randomized controlled experiments because deploying police officers is costly and experiments are time consuming. Klick et al. [8]said in cities (such as Washington, D.C.), this issue is easily addressed. Washington D.C. has a terrorist alert system due to the natural terrorist threat on the capital. The city responds to high-alert days by increasing the number of police officers in specific parts of the city. Assuming terrorist attacks are uncorrelated to crime in the streets, this existing plan to improve security provides a natural randomized experiment for measuring how numbers of officers affect crime. A researcher must observe only how crime rates change in the areas where officers are assigned due to terrorist attacks. In general, to perform a natural experiment, a researcher needs external knowledge about the phenomenon being studied. In our example, this external knowledge is the existence of the terrorist-alert system.

Natural experiments are also effective in social media research. As in controlled experiments, natural experiments can help identify influence in social media. Influence is known to be temporal; that is, the influential individuals take an action or adopt a product before their influencees do. Anagnostopoulos et al. [2] proposed to obtain randomized control groups by shuffling the times users are influenced in social media; for example, individuals who influence others to buy books must have read those books before the others, so by shuffling the times books are recommended and keeping the times when books were purchased intact, researchers can create randomized control groups. Their approach, known as the shuffle test, identifies influence by measuring it on the original social media site and comparing it to its value on these randomized control groups. Likewise, Christakis et al [4] showed that since influence is directional, or the influencer influences the influencee, the researcher can create randomized control groups by reversing the direction of interactions in social media; for instance, assume a researcher is measuring the influence a popular music star has over his or her fans. The researcher can observe and count the number of fans joining some site after the star joins the site. To compute influence for a randomized control, the researcher can reverse the influence direction by counting the number of fans joining the site before the star does.

Natural experiments are difficult to perform in some cases, as they require searching for subtle ways to construct the control group. In general, a researcher might not be able to perform a natural experiment or have the luxury of randomized experiments. In this scenario, the researcher can resort to nonequivalent control.

Table 1: *Methods for Determining True Causality.*

Method	**Control Group**	**Treatment Group**
Controlled Experiments	One or many **manually** randomized control groups	One or many **manually** treated treatment groups
Natural Experiments	Naturally random groups (randomization unnecessary)	**Natural** treatment groups (no treatment required)
Nonequivalent Control	One or many manually **pseudo-randomized** control groups	One or many **manually** treated treatment groups

2.3 Nonequivalent Control

In nonequivalent control, the control group is not selected randomly but such that the control group is similar to a randomized group. Consider user migrations across social media sites. Assume a researcher would like to determine if having no friends on one site will cause a user to migrate to another site. The researcher can collect a random set of users across multiple sites; in it, users with no friends represent the treatment group. To create control groups, the researcher needs random subsets of users with friends. Instead, the researcher can construct similar control groups by taking random samples from the collected users who have friends. For both control and treatment groups, the researcher can observe user activity over time and determine if users migrate. If migrations are more significantly observed in treatment groups than in control groups, the researcher can safely say having no friends is likely to cause migrations. Kumar et al. [11] employed a similar approach to understand user migration patterns in social media.

All methods discussed thus far aim to determine true causality (see the table here). However, at times, determining true causality is impossible, and a researcher can determine only pseudo-causality by employing causality-detection techniques.

2.4 Causality Detection

Assume the researcher would like to determine if the number of friends X_t a specific user has on a site at time t causes the number of posts Y_t the user publishes at that time. Causality-detection methods help validate this hypothesis by finding the relation between the two temporal variables $X = \{X_1, X_2, \dots, X_t, X_{t+1}, \dots\}$ and $Y = \{Y_1, Y_2, \dots, Y_t, Y_{t+1}, \dots\}$. A well-known causality-detection technique is Granger causality, [6, 22], first introduced by Clive W.J. Granger, a Nobel laureate in economics.

Given two variables $X = \{X_1, X_2, \dots, X_t, X_{t+1}, \dots\}$ and $Y = \{Y_1, Y_2, \dots, Y_t, Y_{t+1}, \dots\}$, variable X Granger causes variable Y when historical values of X help better predict Y than just using the historical values of Y (see Figure 2).

Consider a linear-regression model for predicting Y. The researcher can predict Y_{t+1}

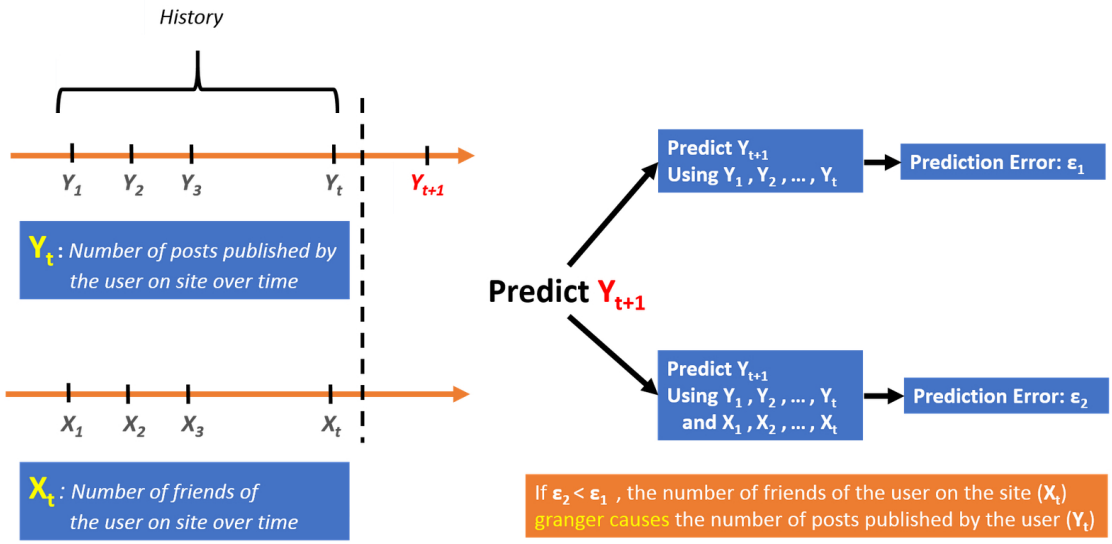

Figure 2: *Granger Causality Example.*

by using either $Y_1, \ldots, Y_t$ or a combination of $X_1, \ldots, X_t$ and $Y_1, \ldots, Y_t$,

$$Y_{t+1} = \sum_{i=1}^{t} a_i Y_i + \epsilon_1, \tag{1}$$

$$Y_{t+1} = \sum_{i=1}^{t} a_i Y_i + \sum_{i=1}^{t} b_i X_i + \epsilon_2, \tag{2}$$

where ϵ_1 and ϵ_2 are the regression model errors. Here, $\epsilon_2 < \epsilon_1$ indicates that using X helps reduce the error. In this case, X Granger causes Y.

Note Granger causality is pseudocausality, not equivalent to the true causality; for instance, a variable Z that causes both X and Y can result in X Granger causing Y. In this case, Granger causality is not equivalent to true causality.

3 Outcome Evaluation

Consider evaluating a method that estimates the virality of rumors in a social network without ground truth. The researcher expects the method to perform well for some user population, though the size of this population is often unknown. The researcher thus needs to take three independent steps:

- *Estimate magnitude.* Estimate the size of the user population to which the rumor can spread. It is imperative to know the size of the population that has the opportunity to observe the rumor, as a rumor that is viral in a small population may not be as viral in a larger population;

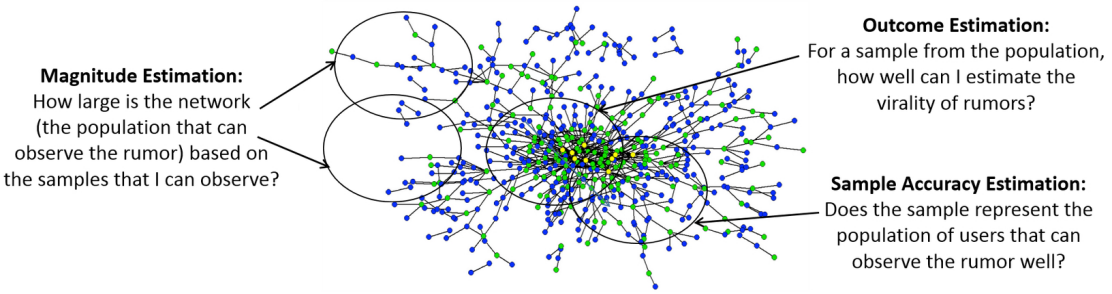

Figure 3: *Here, we are evaluating a method that estimates the virality of rumors in a social network without ground truth. Green nodes are users who have not received the rumor yet. Blue nodes are users who have received the rumor and have adopted (such as by posting) it. Yellow nodes are users who have received the rumor but rejected it. As there is no ground truth, the method that estimates the virality of rumors lacks access to node colors.*

- *Estimate sample accuracy.* Given the population of users that has the opportunity to observe the rumor, the researcher often has limited access or computational power to analyze the whole population and must thus sample the population. This sample must accurately represent the general population. Unfortunately, random sampling is not always possible in the restricted ecosystems of social media; and
- *Estimate outcome.* Over a sample, a researcher must determine how well the virality of rumors is estimated, despite lacking ground truth.

Figure 3 outlines these steps for the method that estimates rumor virality in a social network.

3.1 Magnitude Estimation

For estimating the size of a population without ground truth, a researcher can resort to similar techniques designed in anthropology or ethology. In anthropology, the network scale-up method, or NSUM, introduced by Bernard et al. [18] is designed for magnitude estimation. NSUM was first introduced after the devastating 1985 earthquake in Mexico City to predict, after the earthquake had struck, the number of individuals likely to have perished during the earthquake. To estimate the number of people people likely to have perished, the researcher can ask a small set of individuals if they personally know someone who perished during the earthquake. By surveying a small set of individuals, the researcher can estimate p, the probability of individuals personally knowing someone who perished during the earthquake. Let n denote the size of the city, s the size of the subpopulation that perished, n_i^s the number of perished people personally known by individual i, and f_i the size of the population individual i knows personally (number of

friends of individual i). NSUM predicts s as

$$s = n \times p = n \times \frac{\sum_i n_i^s}{\sum_i f_i}. \tag{3}$$

NSUM and variants [12] have been used successfully in social media research; for instance, researchers can estimate the number of users from a specific country in Facebook (s), using the number of users on Facebook (n), number of friends users have on Facebook (f_i), and number of friends they have from that country (n_i^s).

A similar approach, called mark and recapture, is often used to estimate the population of animals, taking place in two phases: In the first, the researcher captures a set of animals and marks them; in the second, after some preset time, the researcher captures another set of animals and observes how many were recaptured. Let n denote the size of the animal population being estimated, m the total population marked in the first phase, c the population captured in the second phase, and r the population that was marked in phase one and recaptured in phase two. The mark-and-recapture technique then estimates n as

$$n = \frac{m \times c}{r}. \tag{4}$$

As with NSUM, this approach can be used to estimate population sizes in social media, as shown by Papagelis et al. [17].

3.2 Sample Accuracy Estimation

In social media, researchers often sample users or content from a site, despite not knowing how representative the samples actually are. Determining effective sample size is well studied in statistics and in survey sampling [7], when the sampling technique is known. But when the sampling technique is unknown, the researcher can generate a small random sample and compare it to the sample obtained through the unknown technique. This way the researcher can empirically determine how close the unknown samples are to random samples; for instance, Morstatter et al. [15] in 2013 estimated how representative tweet samples collected from Twitter's API are by comparing them to the random samples collected from all the tweets on Twitter during the same period. By comparing samples through statistical measures, they concluded that small samples from Twitter's API might, in some cases, not be representative of the general tweets observed in Twitter.

3.3 Outcome Estimation

When a representative sample is available, evaluating the performance of a computational method on it without ground truth can be a challenge. Depending on whether user feedback or controlled experiments are available from the social media site, different techniques can be employed to estimate outcome.

3.3.1 Unavailable Feedback

When feedback is not available, the researcher can use external sources or computational methods to perform the evaluation; for instance, consider a method that identifies influential users on a site. To verify they are correctly identified, the researcher can seek validating evidence on other sites; for instance, Agarwal et al. [1], verified influential bloggers identified by their algorithm on The Unofficial Apple Weblog (TUAW) using information available on the Digg news-aggregator site. They assumed influential bloggers on TUAW should be praised and cited more often on sites like Digg. As another example of unavailable feedback, consider a method that identifies user communities on a social network. As such communities are often unknown, the researcher can evaluate identified communities through methods that quantify the intuition that a community corresponds to a densely interacting set of users [20]. One such method is modularity [16] which is often used for evaluating community-detection algorithms in social media when ground truth is unavailable [21].

3.3.2 Available Feedback

When feedback is available, the researcher can perform controlled experiments to evaluate. In the industry, these experiments are often referred to as A/B testing;[9, 10]. for example, a social media company tests a new feature for its site by dividing users on the site randomly into two groupsA and B. Group A is shown the feature, and for its users, some quantity (such as number of visits to the site) is measured. Group B is shown the same site but without the feature, and its number of visits to the site is measured as well. If the difference between the number of visits in group A is significantly larger than the number of visits in group B, then the feature is beneficial to the site and can be expected to increase the traffic.

4 Conclusion

Evaluation of social media means new challenges. One is the lack of ground truth. However, proven scientific methods can be borrowed and tweaked for evaluating social mediaresearch findings.

Here, we have discussed three general categories of evaluation on social media: spatiotemporal, causality, and outcome. Spatiotemporal evaluation can be performed knowing that humans exhibit periodical spatiotemporal behavior. Causality evaluation can be performed by investigating the counterfactual, performing controlled, randomized, or natural experiments, finding nonequivalent controls, or employing causality detection techniques. When evaluating outcome, three tasks must be performed: estimating magnitude, estimating sample accuracy, and estimating outcome (see Figure 4). Learning and applying methods from statistics, anthropology, and ethology can help researchers perform these tasks efficiently. In addition, these methods help advance research in social media and inspire development of novel evaluation methods for new research needs.

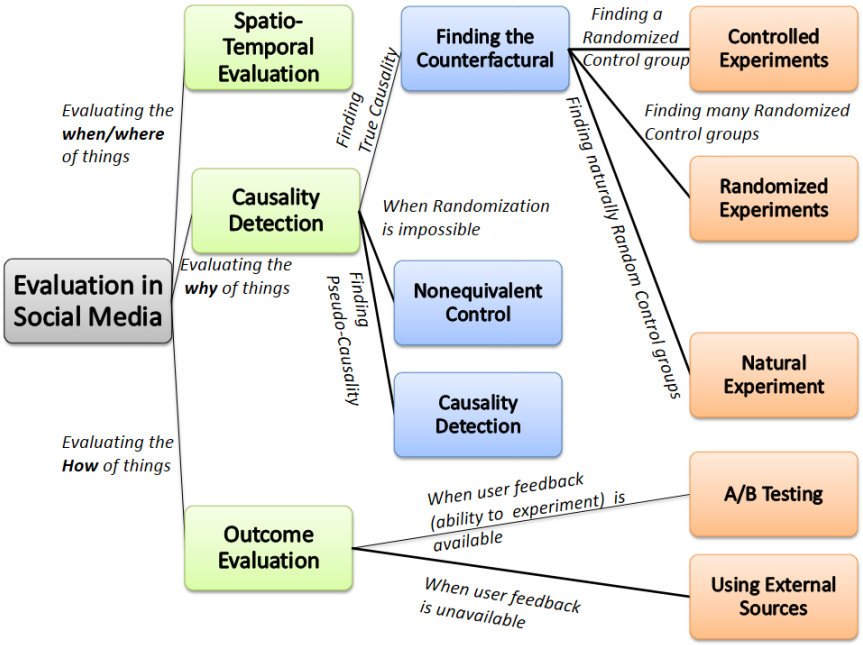

Figure 4: *Map for Evaluation without Ground Truth in Social Media.*

References

[1] Nitin Agarwal, Huan Liu, Lei Tang, and Philip S Yu. Identifying the influential bloggers in a community. In *Proceedings of the 2008 international conference on web search and data mining*, pages 207–218. ACM, 2008.

[2] Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. Influence and correlation in social networks. In *KDD '08*, pages 7–15, New York, NY, USA, 2008.

[3] Geoffrey Barbier, Reza Zafarani, Huiji Gao, Gabriel Fung, and Huan Liu. Maximizing benefits from crowdsourced data. *Computational and Mathematical Organization Theory*, 18(3):257–279, 2012.

[4] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.

[5] Thomas G Dietterich. Ensemble methods in machine learning. In *Multiple classifier systems*, pages 1–15. Springer, 2000.

[6] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[7] Leslie Kish. *Survey Sampling*. New York: Wiley, 1965.

[8] Jonathan Klick and Alexander Tabarrok. Using terror alert levels to estimate the effect of police on crime. In *American Law & Economics Association Annual Meetings*, page 38. bepress, 2004.

[9] Ron Kohavi, Randal M Henne, and Dan Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *Proceedings of the 13th ACM SIGKDD*, pages 959–967. ACM, 2007.

[10] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M Henne. Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*, 18(1):140–181, 2009.

[11] S. Kumar, R. Zafarani, and H. Liu. Understanding User Migration Patterns in Social Media. In *AAAI*, pages 1204–1209, 2011.

[12] Maciej Kurant, Minas Gjoka, Carter T Butts, and Athina Markopoulou. Walking on a graph with a magnifying glass: stratified sampling via weighted random walks. In *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pages 281–292. ACM, 2011.

[13] Timothy La Fond and Jennifer Neville. Randomization tests for distinguishing social influence and homophily effects. In *Proceedings of the 19th WWW conference*, pages 601–610. ACM, 2010.

[14] Winter Mason and Siddharth Suri. Conducting behavioral research on amazons mechanical turk. *Behavior research methods*, 44(1):1–23, 2012.

[15] Fred Morstatter, Jurgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from twitters streaming api with twitters firehose. *Proceedings of ICWSM*, 2013.

[16] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[17] Manos Papagelis, Gautam Das, and Nick Koudas. Sampling online social networks. *Knowledge and Data Engineering, IEEE Transactions on*, 25(3):662–676, 2013.

[18] H Russell Bernard, Eugene C Johnsen, Peter D Killworth, and Scott Robinson. Estimating the size of an average personal network and of an event subpopulation: Some empirical results. *Social science research*, 20(2):109–121, 1991.

[19] P.N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Pearson International Edition. Pearson Addison Wesley, 2006.

[20] Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, page 3. ACM, 2012.

[21] Tianbao Yang, Yun Chi, Shenghuo Zhu, Yihong Gong, and Rong Jin. Detecting communities and their evolutions in dynamic social networksa bayesian approach. *Machine learning*, 82(2):157–189, 2011.

[22] Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu. *Social Media Mining: An Introduction*. Cambridge Univ Press, 2014.