Imputation of Missing Network Data

Chapter · January 2014

CITATION
1

READS
100

1 author:

Mark Huisman
University of Groningen
57 PUBLICATIONS **1,301** CITATIONS

SEE PROFILE

ICTs

▸ Social Networks and Politics

Identify Influential Nodes

▸ Ranking Methods for Networks

Impact

▸ Time- and Event-Driven Modeling of Blogger Influence

Impacts of Policy

▸ Social Media Policy in the Workplace: User Awareness

Importance Ranking

▸ Ranking Methods for Networks

Impression Management

▸ Privacy and Disclosure in a Social Networking Community

Imputation

▸ Imputation of Missing Network Data: Some Simple Procedures

Imputation of Missing Network Data: Some Simple Procedures

Mark Huisman
Department of Sociology/ICS, University of Groningen, Groningen, The Netherlands

Synonyms

Exploratory analyses; Imputation; Link prediction; Missing data; Missing data mechanisms; Non-response; Reconstruction

Glossary

Actor Non-response (Unit Non-response) Missing all outgoing ties of an actor

Tie Non-response (Item Non-response) Missing some ties of an actor

Imputation Substituting missing data by plausible values

Multiple Imputation Repeated stochastic imputation of the same data set after which the results of the analysis are pooled to generate proper estimates of parameters and standard errors

R. Alhajj, J. Rokne (eds.), *Encyclopedia of Social Network Analysis and Mining*,
DOI 10.1007/978-1-4614-6170-8,

MAR Missing at Random
MCAR Missing Completely at Random
MNAR Missing Not at Random

Definition

When confronted with missing data, researchers often want to handle the missing observations by substituting plausible values for the missing scores. This practice of filling in missing items is called *imputation* (e.g., Schafer and Graham 2002). Imputation has several advantages: it is more efficient than analyzing complete cases, it gives the opportunity to use information contained in the observed data in predicting the missing scores, and it allows analysis using standard techniques and software on a complete(d) data set that is the same for all following analyses. The idea of imputation "is both seductive and dangerous," in the words of Dempster and Rubin (1983). "It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can be legitimately handled in this way and situations where standard estimators applied to the real and imputed data have substantial biases" (Dempster and Rubin 1983, p. 8).

The shortcomings of imputation are related to bias and uncertainty. Ad hoc imputations can seriously distort data distributions and relationships, and produce biased estimates. Moreover, in subsequent analyses, predicted scores are treated as observed values which leads to overestimating the sample size and underestimation of uncertainty levels. These considerations should be taken into account when imputing missing data in social network studies, which are specifically concerned with estimating structural relationships between (groups of) ties.

Introduction

In many social network studies, missing data constitute a serious problem. Often, popular software packages can only deal with fully observed network data, while others disregard the missing data or treat the missing observations as nonexisting. These practices result in (serious) loss of information, leading to decreased statistical power, and may lead to serious bias due to the systematic nature of the missingness (e.g., Schafer and Graham 2002; Graham 2009). Moreover, due to the complex dependencies that exist within networks, missing scores of one actor will influence the local neighborhoods of other actors (directly or indirectly via others). This makes careful treatment of missing network data essential.

Missing data treatment procedures that are common in statistical literature can roughly be classified into four categories: analysis of available data, (re)weighting data, likelihood-based procedures (e.g., the EM algorithm), and imputation. Much is already known about the effects of missingness on (statistical) data analysis and the effectiveness of the various treatment procedures (e.g., Schafer and Graham 2002; Graham 2009). However, the effects of missing data treatments on the structural properties of social networks are less often studied (Huisman 2009; Koskinen et al. 2010). In this article, we investigate in which way imputation can be used to treat missing network data. We translate common imputation strategies to the context of social network data and inspect the effect of imputation on network properties.

Graham (2009) recommends that researchers use missing data procedures from the latter two categories: likelihood-based methods and multiple imputation. He calls these methods the "modern" missing data procedures (p. 555). For social network analysis, such "modern" procedures were proposed by Robins et al. (2004), Handcock and Gile (2010), and Koskinen et al. (2010), who describe model-based approaches (likelihood-based or Bayesian) within the framework of exponential random graph models (ERGMs; see Lusher et al. 2013). In the proposed approaches, values for the missing data are simulated in the course of parameter estimation, and observed statistics are replaced by expected values based on these simulations, in a manner similar to the EM algorithm approach (e.g., Schafer and Graham 2002; Graham 2009).

For longitudinal network data, Snijders (2005) proposed a model-based procedure incorporated in stochastic actor-driven models for network evolution. Analogous to the EM algorithm, the model-based procedures can also be used for link prediction and imputation of missing ties.

This paper is concerned with imputation methods. The "modern" imputation procedure that is recommended by Graham (2009) is *multiple imputation* (Rubin 1987). In multiple imputation, each missing value is replaced multiple times (say, m) by random draws from the distribution of the missing values given the observed scores. This results in m completed data sets, which are analyzed separately using the same complete-data method. Finally, the m results are combined such that the final results (i.e., standard errors and p-values) reflect the extra uncertainty due to missing data. Although this procedure is generally recommended as the best way to impute, simple, single imputations can still be useful for specific analyses that do not require hypothesis testing or confidence intervals (Graham 2009). Such analyses are not uncommon for social networks (e.g., blockmodeling). As generating multiple imputations and combining the results of the separate analyses can be a difficult task, single imputation methods are useful to treat missing network data (also as a first step to multiple imputations). These simple imputation methods for missing network data are discussed in this paper.

Key Points

In this paper, we assume a fixed and known set of actors and a single, binary relation between the actors. The tie variable X_{ij} indicates whether the tie from actor i to j is present ($X_{ij} = 1$) or absent ($X_{ij} = 0$). The relation can either be directed, from i to j, or undirected, in which case $X_{ij} = X_{ji}$. Additional information on the ties and/or actors may be available in the form of dyadic covariates or actor-attribute variables. In all three types of variables (tie variables, dyadic covariates, and actor attributes), missing values may occur.

We only consider the situation where missing data is caused by non-response (see Kossinets 2006, or Žnidaršič et al. 2012b, for other sources of sampling errors and missing data in the context of social networks) and distinguish two types of non-response: *unit non-response*, where all scores of an actor are missing (ties and attributes), and *item non-response*, where only particular items are missing. When only tie variables (i.e., network data) are concerned, these two types are also called *actor non-response* and *tie non-response*, respectively (Huisman 2009; Žnidaršič et al. 2012a). A special case of item non-response may occur when all outgoing ties of an actor are missing (actor non-response), but attribute information is available or vice versa. This form of non-response is sometimes called *partial non-response* (De Leeuw et al. 2003). A specific form of partial non-response is common in longitudinal studies, *wave non-response*, which arises when complete network information for an actor is missing at one (or more) measurement moments (see Huisman and Steglich 2008).

The type of non-response determines the amount of data that is available for each actor. With actor non-response, more information is missing for each actor than with partial or item non-response. An advantage of social network data is that information on the network context of incompletely observed actors is often available, at least partially, through observed nominations by other actors. This information can be used to analyze (or even "reconstruct") the network neighborhood of missing actors and should not be omitted from the analyses. This approach supposes that the observed and missing data are not systematically different and that all necessary information about the missing data can be found the observed data. In the statistical literature, this situation is known as data that are *Missing at Random* (MAR; Rubin 1976). When data are MAR, the probability of missingness is related to the observed data, and not to the missing data. If, in addition, the missingness is not related to the observed data either, the data are called *Missing Completely at Random*. If, on the other hand, the probability of missingness is related to the missing (and therefore unknown) values

themselves, the data are *Missing Not at Random* (MNAR). Huisman (2009) provides more details on missing data mechanisms for social network data, and Handcock and Gile (2010) give formal definitions.

The "modern" missing data methods of Graham (2009) assume MAR. This means that all information about the missingness is contained in the observed data, and given these data the missing data mechanism is ignorable. In this situation, the causes of missingness do not have to be taken into account (Koskinen et al. 2010). Simple (older) missing data methods only give unbiased results when data are MCAR, which is only realistic when there is no reason to assume that actors differ in their propensity to fill in network questionnaires (Huisman and Steglich 2008).

Historical Background

One of the first studies on the effects of non-response on the structural properties of social networks is the study by Burt (1987). In this article, he calls missing data "doubly a curse to survey network analysis" (p. 63), because the complexity of network questionnaires is more likely to generate missing data, and the dependence structure of the network increases the impact of missing ties. Others followed up on this study and found, among others, that missing data have a negative effect on network mapping (Borgatti and Molina 2003), underestimate the strength of relationships (Burt 1987), make centrality measures and degree measures unstable (Costenbader and Valente 2003; Kossinets 2006; Borgatti et al. 2006; Huisman 2009), underestimate clustering coefficients (Kossinets 2006; Huisman 2009), and underestimate reciprocity measures (especially in directed graphs; Huisman 2009).

Some of these studies showed that the extent to which structural properties of the network are affected by missing ties also depends on how the available information is used to calculate the measures. For instance, measures based on indegrees are reasonably robust to small proportions of missing data when the observed incoming ties of nonrespondents are included in the analyses (Costenbader and Valente 2003). The same was found for reciprocity measures (Huisman 2009), especially in undirected graphs. This is the result of the unique property of social networks that nonparticipation (or partial participation) does not necessarily mean that the missing actors are not included in the analyses (Borgatti and Molina 2003), that is, when incoming ties of respondents to nonrespondents are available. Because of this property, Stork and Richards (1992) proposed using the information in partially described ties of nonrespondents to reconstruct the missing part of the network.

Stork and Richards (1992) explore problems in analyzing incomplete network data due to non-response. They propose an analysis method for incomplete data based on reconstruction of the missing ties, and make suggestions for designing network studies that improve response rates and that provide information to make decisions about analysis methods. The impact of non-response on network properties was further explored by Kossinets (2006). He investigated a broader set of missing data sources, including boundary specification problems, non-response, and fixed choice designs. An even broader set of sources of measurement error in network data is discussed by Butts (2003), Žnidaršič et al. (2012a), and Wang et al. (2012).

Imputation of Missing Network Data

Imputation is a general and popular approach to handle missing data, and various imputation procedures are thoroughly studied in the statistical literature. Schafer and Graham (2002) give a general overview by distinguishing four classes of single-imputation methods. Before discussing imputation methods for network data, we first present these general classes.

Imputing Unconditional Means A simple (ad hoc) procedure is replacing each missing value with the mean over the observed cases of that item. Although the means of items are preserved, variances and covariances (relations)

are often severely biased. Rounding the mean values, in case of binary or categorical data, even adds more error to the imputed values. Although the added variability is random, it is better to keep rounding to a minimum (Graham 2009).

Imputing from Unconditional Distributions The underestimation of variances by imputing means can be corrected by using the observed (empirical) distributions of the items to impute the missing scores. One class of procedures, called hot-deck procedures, simulates these distributions by (randomly) selecting an observed donor case from the same data set, and missing values are replaced with the observed scores of the donor (e.g., Sande 1982). Although such procedures preserve univariate distributions of variables (i.e., means and variances), relations are still biased.

Imputing Conditional Means Prediction of mean values can be improved using a formal model that accurately captures the association between a missing item and observed items. Often linear (regression) models are used to predict the conditional means of the missing items. These procedures result in more accurate predictions of the missing scores and yield unbiased estimates of means, but underestimate variances and generally overestimate covariances.

Imputing from Conditional Distributions The biases in variances and covariances found in the previous procedures are largely reduced by using conditional distributions to impute the missing values, conditional on observed variables. The conditional distribution of the missing values is simulated using the imputation models of the previous procedure (imputing conditional means), conditional on the observed independent variables in the model. The missing scores are replaced by draws from this distribution. In the practice of empirical research, this procedure usually amounts to imputing regression predictions with an added error term, randomly drawn from the normal distribution (of which the standard error is estimated in the regression analysis; for this purpose, a t-distribution is also often used instead of the normal distribution). Multiple imputation procedures fall in this class of imputations methods (Schafer and Graham 2002; Graham 2009).

Imputation of Missing Ties

Based on the classification of Schafer and Graham (2002), Huisman (2009) presents an overview of simple imputation methods to impute missing ties caused by both actor and tie non-response, which were previously applied in empirical research. These methods belong to the first two classes of imputation methods, and some of them were also investigated by others (Ouzienko and Obradovic 2011; Žnidaršič et al. 2012a). In this section, simple imputation procedures for missing network data will be presented together with more sophisticated imputation models that are more recently proposed in the literature.

Imputing Unconditional Means

For binary tie variables, the total mean value equals the network density. Rounding the density (using a threshold of 0.5) results in filling in zeros in sparse networks and ones in dense networks. In the former case, missing ties are treated as absent. This is called *null tie imputation* by Žnidaršič et al. (2012a) and is even sometimes used in dense networks.

Instead of filling in the overall mean value of the tie variable, the mean of the incoming ties of an actor (the "item mean") or the mean of the outgoing ties of an actor (the "person mean") can be used. The latter method can obviously not be used in the case of actor non-response. The former method results in imputing the modal value of the incoming ties and is called *imputation based on Modal indegree values* by Žnidaršič et al. (2012a).

Reconstruction

Stork and Richards (1992) suggest reconstructing the missing part of the network by replacing the missing outgoing ties of nonrespondents by observed incoming ties to these actors. As a result, that part of the network with ties between respondents and nonrespondents becomes symmetric.

Additional imputations are required for ties between nonrespondents. For these, Huisman and Steglich (2008) and Huisman (2009) use random imputation proportional to observed density (i.e., the probability of tie is equal to the observed density of the network). Žnidaršič et al. (2012a) propose additional imputations based on modal indegree values.

Note that reconstruction is an imputation procedure when applied to directed networks. For undirected networks it is an available-case method using partially described links (i.e., reported by only one of the two actors; Stork and Richards 1992), because no new ties are added. The underlying assumption is that the tie between two actors can be measured by the report of only one of the actors and that respondents and nonrespondents do not systematically differ in reporting their relationship. In directed networks, the two tie variables X_{ij} and X_{ji} are allowed to differ (in asymmetric dyads), and reconstruction is an imputation method in which missing ties are replaced with plausible values: the reversed tie within the dyad.

Preferential Attachment

Preferential attachment was proposed by Barabasi and Albert (1999) as a model for the growth of networks and was used by Huisman and Steglich (2008) as an imputation model. The model states that the probability that a new tie $X_{ij} = 1$ will emerge between actors i and j is proportional to the current number of neighbors (i.e., indegree) of actor j. This means that the probability that an actor (observed or missing) will link to another is dependent on the connectivity of others. Liben-Nowell and Kleinberg (2007) mention preferential attachment as a method for link prediction. Huisman and Steglich (2008) propose a two-step imputation procedure based on random draws from outdegree distributions and random draws using preferential attachment probabilities to impute missing data caused by actor non-response. Huisman (2009) investigates this method also for tie non-response.

Hot Deck Imputation

Hot deck imputation uses completely observed donor actors to replace all ties of the missing actor (actor non-response), or the missing ties of an incomplete actor (tie non-response). Donor actors can be selected using observed attribute values or structural properties of the network, or both. Huisman (2009) gives an example of the latter option and describes a procedure where actors are matched on the attribute alcohol consumption (in a friendship network) and on indegrees. Instead of finding one donor actor (the "best" donor), a set of donors can be selected, from which one is randomly chosen. Note that reconstruction, as discussed in the previous section, can be regarded as hot deck imputation, defining the donor actor as the second actor in the dyad whose incoming tie is not observed.

Link Prediction

The simple methods presented above all have in common that they are not model based. Although some depend on (often strong) network properties like reciprocity (i.e., reconstruction) or connectivity (i.e., preferential attachment based on indegrees), they do not use a statistical model for network data to relate observed and unobserved scores. Imputation methods that use such models (the conditional methods, in the classification of Schafer and Graham 2002) are the link-prediction methods based on stochastic blockmodels proposed by Guimerà and Sales-Pardo (2009); methods based on latent factor models proposed by Hoff (2009); methods based on ERGMs proposed by Koskinen et al. (2010) and Handcock and Gile (2010); and methods based on Kronecker graph models proposed by Kim and Leskovec (2011). For longitudinal network data, Snijders (2005) proposed a model-based procedure incorporated in stochastic actor-driven models for network evolution. This procedure is described and investigated by Huisman and Steglich (2008), who call it a *hybrid imputation method* as it only uses initial imputations at the first observation moment and does not automatically result in a completed data set.

Imputation of Missing Actor Attributes

Missing actor attributes could be regarded as "ordinary" missing data in any non-network data set and treated separately from the network data. As discussed previously, ample imputation methods are available and are well known and discussed in statistical literature. For example, a number of completed attribute sets can be created with the "modern" missing data technique of multiple imputation (Schafer and Graham 2002; Graham 2009). These sets are then used in subsequent network analyses, after which the results are pooled. Ouzienko and Obradovic (2011) present some simple imputation methods (e.g., mean imputation), to impute the missing attributes without taking into account the tie variables.

Actor attributes, however, are known to be (often strongly) related to structural properties of the network. Imputation of missing attribute data should therefore be considered within a general framework of imputations together with missing tie variables. As Graham puts it, "all variables in the analyses model must be included in the imputation model" (p. 559). Omitting variables from the imputation model amounts to assuming that there is no association between the omitted variables and the included variables. This may lead to underestimation of relations, that is, the structural properties of the network. Koskinen et al. (2009) illustrate the added difficulty of missing attributes in a model-based (i.e., ERGM) framework, but practical solutions for the combined imputation of missing attribute and network data are not yet available.

Key Applications

Huisman and Steglich (2008) investigated the effect of imputation of missing ties in longitudinal network data suffering from wave non-response. They compared two simple methods (reconstruction and imputation based on preferential attachment) and the model-based procedure of Snijders (2005), and studied the effects on parameter estimates for actor-driven models for network evolution. They found that the latter, model-based method generally had smaller biases than the simple imputations. This is as expected, given that the simple imputation methods do not take into account the longitudinal aspect of the data, whereas the model-based method was designed to do so.

Huisman (2009) investigated simple methods (all simple methods mentioned in the previous section) for cross-sectional network data suffering from actor non-response and tie non-response. He examined the effects of the methods on some structural properties of social networks (e.g., reciprocity, clustering, assortativity). A similarly designed study was performed by Žnidaršič et al. (2012a, b), who investigated the effect of simple imputations on blockmodeling. All studies arrive at the same general conclusion that the majority of simple imputation procedures can severely bias estimates of network properties. In such cases, a complete-case analysis is to be preferred. Still, specific situations have been identified in which the simple methods give good results. Most noteworthy in this respect is the reconstruction method, which performs reasonably well in cases where reciprocity is high and the proportion missing data is low.

Future Directions

It was found that simple imputation methods are unable to capture the structural properties of networks, because relationships are not incorporated in the imputation models and therefore poorly estimated. The simple methods are not model based and do not clearly specify how the observed features of the data relate to the missing data. Only very simple structures like density (degrees) and reciprocity are partially taken into account (by preferential attachment and reconstruction, respectively), and for these properties, some simple methods can give satisfactory results, especially the reconstruction method based on the important mechanism of reciprocity observed in many social networks. Therefore, reconstruction is a method we should consider to improve in

future studies. Multiple imputation is a procedure we certainly want to explore in this respect, as repeated imputations are needed in "modern" missing data methods to give correct estimates of uncertainty levels needed for inferences.

In this paper, actor attributes played a minor role, both as predictors in imputation models and as (missing) response variables to be predicted by the imputation models. We would like to explore the possibilities to use attributes in imputation models and methods for the combined imputation of attributes and network data. As stated above, practical solutions with respect to imputation of actor attributes are topics of ongoing research.

Cross-References

- ▸ Exponential Random Graph Models
- ▸ Link Prediction: A Primer
- ▸ Research Designs for Social Network Analysis
- ▸ Sampling Effects in Social Network Analysis

References

Barabasi A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512

Borgatti SP, Molina JL (2003) Ethical and strategic issues in organizational social network analysis. J Appl Behav Sci 39:337–349

Borgatti SP, Carley KM, Krackhardt D (2006) On the robustness of centrality measures under conditions of imperfect data. Soc Netw 28:124–136

Burt RS (1987) A note on missing network data in the general social survey. Soc Netw 9:63–73

Butts CT (2003) Network inference, error, and informant (in)accuracy: a Bayesian approach. Soc Netw 25: 103–140

Costenbader E, Valente TW (2003) The stability of centrality measures when networks are sampled. Soc Netw 25:283–307

De Leeuw ED, Hox JJ, Huisman M (2003) Prevention and treatment of item nonresponse. J Off Stat 19:153–176

Dempster AP, Rubin DB (1983) Overview. In: Madow WG, Olkin I, Rubin DB (eds) Incomplete data in sample surveys vol II: theory and bibliographies. Academic, New York, pp 3–10

Graham JW (2009) Missing data analysis: making it work in the real world. Ann Rev Psychol 60:549–576

Guimerà R, Sales-Pardo M (2009) Missing and spurious interactions and the reconstruction of complex networks. Proc Natl Acad Sci 106:22073–22078

Handcock MS, Gile KJ (2010) Modeling social networks from sampled data. Ann Appl Stat 4:5–25

Hoff P (2009) Multiplicative latent factor models for description and prediction of social networks. Comput Math Organ Theory 15:261–272

Huisman M (2009) Imputation of missing network data: some simple procedures. J Soc Struct 10:1–29

Huisman M, Steglich CEG (2008) Treatment of non-response in longitudinal network studies. Soc Netw 30:297–308

Kim M, Leskovec J (2011) The network completion problem: inferring missing nodes and edges in networks. In: SIAM international conference on data mining (SDM), Mesa, pp 47–58

Koskinen JH, Robins GL, Pattison PE (2009) Missing data in social networks: problems and prospects for model-based inference. MelNet Social Network Laboratory technical report 08-03, Department of Psychology, School of Behavioural Science, University of Melbourne

Koskinen JH, Robins GL, Pattison PE (2010) Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation. Stat Method 7:366–384

Kossinets G (2006) Effects of missing data in social networks. Soc Netw 28:247–268

Liben-Nowell D, Kleinberg J (2007). The link-prediction problem for social networks. J Am Soc Inf Sci Technol 58:1019–1031

Lusher D, Koskinen JH, Robins GL (eds) (2013) Exponential random graphs models for social networks. Structural analysis in the social sciences, vol 35. Cambridge University Press, Cambridge

Ouzienko V, Obradovic Z (2011) Imputation of missing links and attributes in longitudinal social surveys. In: IEEE international conference on data mining workshops, Vancouver, pp 957–964

Robins G, Pattison P, Woolcock J (2004) Missing data in networks: exponential random graph ($p*$) models for networks with non-respondents. Soc Netw 26: 257–283

Rubin DB (1976) Inference and missing data. Biometrika 63:581–592

Rubin DB (1987) Multiple imputation for nonresponse in surveys. Wiley, New York

Sande IG (1982) Imputation in surveys: coping with reality. Am Stat 36:145–152

Schafer JL, Graham JW (2002) Missing data: our view of the state of the art. Psychol Methods 7:147–177

Snijders TAB (2005) Models for longitudinal network data. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, Cambridge, pp 215–247

Stork D, Richards WD (1992) Nonrespondents in communication network studies. Group Organ Manage 17:193–209

Wang DJ, Shi X, McFarland DA, Leskovec J (2012) Measurement error in network data: a re-classification. Soc Netw 34:396–409

Žnidaršič A, Doreian P, Ferligoj A (2012a) Absent ties in social networks, their treatments, and blockmodeling outcomes. Metodološki Zvezki 9:119–138
Žnidaršič A, Ferligoj A, Doreian P (2012b) Non-response in social networks: the impact of different non-response treatments on the stability of blockmodels. Soc Netw 34:438–450

Incentives in Collaborative Applications

Elina Yaakobovich[1], Rami Puzis[2], and Yuval Elovici[1]
[1]Department of Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[2]University of Maryland Institute for Advanced Computer Studies (UMIACS), College Park, MD, USA

Synonyms

Cooperation; Game theory; Over-justification; Social interaction; The prisoner's dilemma

Glossary

Collaborative System A system that relies on the collaboration of its users when providing a service
Contribution Any user activity that is or might be of benefit to other users or the system itself
Extrinsic Interest Interest that is based on an outside stimulation (such as a reward)
Free Rider A participant who enjoys the contribution of others but fails to contribute or contributes significantly less than others
Intrinsic Interest Interest that is based on inner satisfaction

Definition

Collaborative systems rely on the cooperation of their users to provide a service. We can distinguish between various types of collaborative systems, such as Q&A services, social networking services, sharing platforms, and crowdsourcing platforms. Despite the fact that each one of these services and platforms has its own unique characteristics, they all rely on the cooperation of their participants and community members in order to operate successfully.

The study of incentives and cooperation is multidisciplinary and is approached through diverse and sometimes contradictory points of view. Social science, economy (game theory), and biology are merely a part of the different approaches which attempt to explain cooperation in platforms that require it.

Game theory approaches assume that users are rational and act in order to maximize their benefit (Kreps 1990). These approaches model systems using well-defined game settings and provide mathematically proven strategies for maximizing self and/or overall benefit. However, it was shown that human decision making is biased when operating under risk and uncertainty (Kahneman and Tversky 1979). This bias is not captured by game theoretic approaches.

Moreover, different users have diverse reactions to various types of incentives. For example, pro-self-oriented users will respond best to extrinsic incentives that ensure benefit from using the system. Pro-social-oriented users will better cooperate in communities with social ties and trust.

It is important to understand the population at hand and the system objectives in order to design the most appropriate incentive mechanisms for each collaborative system.

Introduction

Collaborative systems rely on the cooperation of their users to provide service in the form of knowledge or information, products, paid or unpaid tasks, etc. Some well-known examples of collaborative systems, protocols, or applications include informative websites, such as Wikipedia or Yahoo! Answers; crowdsourcing platforms, such as Amazon Mechanical Turk; peer-to-peer file sharing protocols, such as BitTorrent; social