A Credibility Measurement Method of Smart Grid Data

Xiaorong Cheng[1] ,Tianqi Li[1]
[1]North China Electric Power University, Hebei Baoding, China
xiaor_cheng@163.com, ltqwhy@163.com

***Abstract*—The requirement of the credibility is analyzed for smart grid data and the traditional theories of data measurement and bad data identification are studied in power system. On the basis of that, this paper mainly studies the credibility between data sources, the construction of the trusted network generation algorithm, and the hierarchical and dynamic trusted network model innovatively which is based on the relationship between data sources. The credibility evaluation of data sources, the credibility evaluation between data sources and the credibility evaluation of data are studied. The credibility between data sources is restricted by the credibility of data sources, the credibility of data sources is restricted both from the credibility of data and the credibility between data sources, the credibility of data is restricted by the credibility between data sources and the credibility of data sources, they are interrelated and restricted by each other and constitute to a whole. The simulation results show that the model can satisfy the requirement of the credibility of smart grid data better, and provide the ideas to solve the problem of the credibility measure for prospective research is feasible.**

Keywords—Smart Grid; Big Data; Credibility; Dynamic; Trusted Computing; Modeling and Simulation;

I. INTRODUCTION

At present, big data runs through the electricity generation, transmission, electricity distribution, electricity consumption, scheduling and other aspects of the production and management in the smart grid system. It has typical "3V"(Volume, Velocity, Variety) and "3E"(Energy, Exchange, Empathy) characteristics. It is not difficult to find the typical "HDC" attribute of smart grid big data, among which are Heterogeneity, Dynamic and Complexity. Therefore, the electricity data sets must be filled with a large number of unreliable data, which will reduce the convergence performance of state estimation of the power system, affecting the dispatcher to make the wrong decisions, thereby resulting in the abnormal operation of the smart grid system, and may even threaten the safety of the electric power system. If we can evaluate the credibility of the original data, we can effectively reduce the risks and improve the credibility of big data.

To build a strong and smart grid, people have a higher requirement for the credibility of data and the reliability of network operation, which is urgent to study the trusted measurements and evaluation methods of big data. Therefore, the paper proposes a new model of credibility measurement for smart grid data.

978-1-4673-9613-4/16/$31.00 ©2016 IEEE

II. RESEARCH ON THE DATA CREDIBILITY IN SMART GRID

A. Research on bad data identification

So far, bad data identification algorithms are mainly based on state estimation and data mining in the power system. On the one hand, State estimation is to improve the accuracy of the data by using the redundancy of real-time measurement system, and to eliminate the error caused by random interference, and then estimate the running state of the system. These methods mentioned in the literature [3-4] are prone to the residual pollution and residual error, resulting in missing or false detection, especially when there are multiple bad data, these methods often occurs the phenomenon of false identification.

B. Research on data credibility

At present, there are many research methods and some results for the trusted measurement and evaluation of data. The methods of credibility analysis are mainly divided into two categories, one is subjective trust analysis based on belief, which is a cognitive phenomenon which is the subjective judgment of the specific characteristics or behavior of the object of trust, and this kind of judgment, which has ambiguity, uncertainty and can't be accurately described, verified and speculated, is relatively independent of the subject's characteristics and behavior [5]. Documents [5-6] have proposed different subjective methods based on Probability, Fuzzy Set Theory, Cloud Theory and so on. The other is the objective trust analysis based on the evidence theory, which can be accurately described, verified and speculated. The trust relationship between the two is strictly defined by appropriate evidence. D-S evidence theory is used to calculate the credibility by documents [8-9].

Although the models take the dynamic interaction and randomness between the entities into account, it can not consider the impact of timeliness and malicious recommendation, and lack of flexibility. Once the weight is determined, the system is very difficult to adjust it, it will result in a lack of adaptive prediction model. Therefore, it is urgent to study the problems of the credibility measure and service of big data.

III. TRUSTED ANALYSIS MODEL FOR SMART GRID DATA

Aiming at the characteristics of big data's "3V", "3E" and "HDC", this paper presents a model of big data's credibility measure of dynamic construction, which is divided into three parts - the trusted measurement model between data sources, the trusted measurement model of data sources and the trusted measurement model of data. The credibility between data sources is restricted by the credibility of data sources, the credibility of the data sources is restricted by the credibility of data and the credibility of the data source, the credibility of data is restricted by the credibility between data sources and the credibility of data sources, they are interrelated and restricted each other and constitute a whole.

A. Related notion

In order to understand the model, the relevant definitions of the method are given in this paper, which are used to explain the basic issues in the analysis of smart grid big data's credibility.

Data source: It refers to the provider of data in the smart grid's environment.

Data: It refers to the characteristics of multiple attributes. Its notation is denoted as data=$\{d_1,d_2,d_3,\ldots,d_n\}$. Thereinto, d_i refers to the "i" attribute of the data.

Trusted network: It refers to a network composed of data sources and directed links between them.

B. A trusted measurement model between data sources

In the process of calculating the credibility of the smart grid, when there is a direct context interaction among data sources or the similarity of data or behaviors provided by between data sources exceeds a certain threshold. The direct link of the directed graph can be established among data sources. With the expansion of the network size, the trusted network is becoming more and more stable. If it finds out that a data source is not trusted, the model can also quickly impose the penalty factor on the credibility of the provider (data source), making it less reliable for the provider for a period of time. But as time goes on, if the data source can continue to provide reliable data, its credibility will be restored. If data sources have no new context in a calculation interval in the network model of the credibility analysis, time penalty is imposed on them.

Definition 1. **Credibility between data sources:** It is composed of the local credibility and the global credibility between data sources. Its notation is denoted as $Trust_A(B,t)$, the meaning behind which is the comprehensive credibility of data source A relative to data source B at the "t" moment, as is shown below in formula (1).

$$\mathrm{Trust}_A(B,t)=\alpha_1\cdot LocalTrust_A(B,t)+\beta_1\cdot GlobalTrust_A(B,t) \tag{1}$$

Thereinto, $\alpha_1+\beta_1=1$.

Definition 2. **Local credibility:** When there is a direct context interaction between data sources or the similarity of data or behaviors provided by between data sources exceeds a certain threshold, we believe that between data sources have a local credibility. It is composed of the credibility of direct context interaction and the credibility of the similarity between data sources. Its notation is denoted as $LocalTrust_A(B,t)$, the meaning behind which is the local credibility of data source A relative to data source B at the "t" moment, as is shown below in formula (2).

$$LocalTrust_A(B,t)=\begin{cases} Random()\ or\ 0 & ,t=0 \\ LocalTrust_A(B,t-1)\cdot\mu_L(t) & ,\Delta Context(A,B,t)=0 \\ \left[\alpha_2\cdot DirTrust(A,B,Context(A,B,t),t)+\beta_2\cdot Accept(A,B,t)\right]\cdot\lambda_L(t) & ,\ \text{other} \end{cases} \tag{2}$$

Notes:

a) The initial value is a random number or 0, which indicates that data source A has some trust or no trust for data source B.

b) $\mu_L(t)$ is the time decay factor at the "t" moment. If the local credibility of data source A is the same as that of data source B at the t and t-1 moment, then it is punished by the time decay factor. Thereinto, $\mu_L(t)=1-\dfrac{\Delta t}{t-t_0}$, $0\le\mu_L(t)\le 1$.

Δt is the time difference of calculation between two times. t_0 is the starting moment of the current calculation, t is the current moment.

c) $\Delta Context(A,B,t)$ is whether between data source A and data source B has a new context of direct interaction at the "t" moment.

$$\Delta Context(A,B,t)=Context(A,B,t)-Context(A,B,t-1).$$

d) $DirTrust(A,B,Context(A,B,t),t)$ is the trusted value of data source A relative to data source B in the circumstances of context interaction at the "t" moment.

e) $Accept(A,B,t)$ is the recognition value of the similarity of data source A relative to data source B at the "t" moment.

$$Accept(A,B,t)=\frac{\sum_{\substack{data_a\in Data(A)\\ data_b\in Data(B)}} Sim(data_a,data_b)}{Data(A)\cap Data(B)}.$$

$Data(A)$ is a data set provided by data source A. $data$ is a data provided by data source. $Sim(data_a,data_b)$ is the similarity degree between $data_a$ and $data_b$.

$Data(A)\cap Data(B)$ is the number of the same theme in the data sets provided by data source A and B.

f) $\lambda_L(t)$ is the penalty coefficient of local credibility of the model at the "t" moment.

$$\lambda_L(t)=\begin{cases} 1 & ,\Delta LocalTrust_A(B,t)\ge 0 \\ 0\le x<1 & ,\Delta LocalTrust_A(B,t)<0 \end{cases}.$$

$\Delta LocalTrust_A(B,t)$ is whether the local credibility of data source A relative to data source B has changed at the "t" moment.

$$\Delta LocalTrust_A(B,t)=LocalTrust_A(B,t)-LocalTrust_A(B,t-1).$$

g) $\alpha_2+\beta_2=1$.

Definition 3. Global credibility: It refers to the credibility of data source in the trusted network, that is, the credibility of data source. Its notation is denoted as $GlobalTrust_A(B,t)$, the meaning behind which is the global credibility of data source A relative to data source B at the "t" moment, as is shown below in formula (3).

$$GlobalTrust_A(B,t) = Trust(B,t) \tag{3}$$

C. A trusted measurement model of data source

Definition 4. Credibility of data source: It is composed of the expected value of the credibility of all historical data provided by data source and the recommendation credibility of data sources of each layer in the whole trusted network. Its notation is denoted as $Trust(A,t)$, the meaning behind which is the credibility of data source A at the "t" moment, as is shown below in formula (4).

$$Trust(A,t) = \begin{cases} Random()\ or\ 0 & ,t = 0 \\ Trust(A,t-1)\cdot \mu_S(t) & ,\Delta Context(A,B,t) = 0 \\ \left[\alpha_3 \cdot \frac{\sum_{data_a \in Data(A)} Trust(data_a,t)}{Sum(Data(A))} + \beta_3 \cdot (\gamma_n \cdot \mathrm{Re}commend_n(A,t))\right] \cdot \lambda_S(t) & ,\text{other} \end{cases} \tag{4}$$

Notes:

a) The initial value is a random number or 0.

b) $\mu_S(t)$ is the time decay factor at the "t" moment. If the credibility of data source A is the same at the t and t-1 moment, then it is punished by the time decay factor.

$$\mu_S(t) = 1 - \frac{\Delta t}{t - t_0},\ 0 \le \mu_S(t) \le 1.$$

c) $\lambda_S(t)$ is the penalty coefficient of the credibility of data source at the "t" moment.

$$\lambda_S(t) = \begin{cases} 1 & ,\Delta Trust(A,t) \ge 0 \\ 0 \le x < 1 & ,\Delta Trust(A,t) < 0 \end{cases}.$$

Thereinto, $\Delta Trust(A,t)$ is the difference of calculation for data source A at the t and t-1 moment.

$$\Delta Trust(A,t) = Trust(A,t) - Trust(A,t-1).$$

d) $Trust(data_a,t)$ is the credibility of $data_a$.

e) $Sum(Data(A))$ is the total number of data provided by data source A.

f) γ_n is a $1*n$ dimensional vector which consists of trusted weight of every layer relative to the objective data source in the trusted network.

g) $\alpha_3 + \beta_3 = 1$.

Definition 5. Recommendation credibility: It refers to the credibility of data source rSelative to the best path to the objective data source. Its notation is denoted as $\mathrm{Re}commend(A,B,t)$, the meaning behind which is the recommendation credibility of data source A relative to the best path to data source B at the "t" moment.

$\mathrm{Re}commend_n(A,t)$ is the recommendation credibility of each layer of data sources relative to the objective data source A. Thereinto, it is a n*1 dimensional vector, the first element of which is the expected value of recommendation credibility of all data sources of the first layer, and the like, each vector element is the expected value for the corresponding layer. The average number of layer is set according to the accuracy and needs, the greater the number of layer is, the greater the amount of calculation is, and the credibility of the corresponding data is more accurate.

The recommendation credibility of a data source relative to data source A for the "i" layer of trusted network is calculated by the formula (5), as is shown below in formula (5).

$$\mathrm{Re}commend(X_i,A,t) = Trust(X_i,t)\cdot Trust_{X_i}(Neighbor^{\max}(X_i -> A),t) \tag{5}$$

Thereinto, X_i is a data source X of the "i" layer. $Neighbor^{\max}(X_i -> A)$ is a data source with the largest credibility adjacent to X_i on the "i-1" layer.

The expected value of the recommendation credibility of data sources relative to data source A for the "i" layer is calculated by the formula, as is shown below in formula (6).

$$\mathrm{Re}commend(A,t)_{(i)} = \frac{\sum_{X \in Circle_i(A)} \mathrm{Re}commend(X,A,t)}{Sum(Circle_i(A))} \tag{6}$$

Thereinto, $Circle_i(A)$ is a set of all data sources on the "i" layer in the trusted network. $Sum(Circle_i(A))$ is the number of all data sources on the "i" layer.

D. A trusted measurement model of data

Definition 6. Credibility of data: It refers to the probability of complementary events of this unreliable data provided by all the data sources which are direct or related providers in the historical records. Its notation is denoted as $Trust(data,t)$, the meaning behind which is the credibility of data in the whole trusted network at the "t" moment, as is shown below in formula (7).

$$Trust(data,t) = 1 - \prod_{data \in Data(X)} (1 - Trust(X,data,t)) \tag{7}$$

Definition 7. The true credibility of data provided by a data source: It refers to the comprehensive of the direct and indirect credibility of data provided by a data source. Its notation is denoted as $Trust(A,data,t)$, the meaning behind which is the true credibility of data provided by data source A at the "t" moment, as is shown below in formula (8).

$$Trust(A,data,t) = \alpha_4 \cdot DirTrust(A,data,t) + \beta_4 \cdot InDirTrust(A,data,t) \tag{8}$$

Thereinto, $\alpha_4 + \beta_4 = 1$.

Definition 8. Direct credibility of data provided by a data source : It refers to the credibility of data source in the entire trusted network. Its notation is denoted as $DirTrust(A,data,t)$, the meaning behind which is the direct credibility of data provided by data source A at the "t" moment, as is shown below in formula (9).

$$DirTrust(A,data,t) = Trust(A,t) \tag{9}$$

Definition 9. Indirect credibility of data provided by a data source : It refers to the credibility of this data recommended by adjacent data sources with high credibility. Its notation is denoted as $InDirTrust(A,data,t)$, the meaning behind which is the indirect credibility of data recommended by data sources associated with the data source A at the "t" moment, as is shown below in formula (10).

$$InDirTrust(A,data,t) = \frac{\sum_{X \in Neighbor_n(A)} Trust(A,X,t) \cdot Trust(X,data,t)}{n} \tag{10}$$

Thereinto, $Neighbor_n(A)$ is n data sources with high credibility adjacent to data source A.

From the above definition, the relationship is shown in Figure 1.

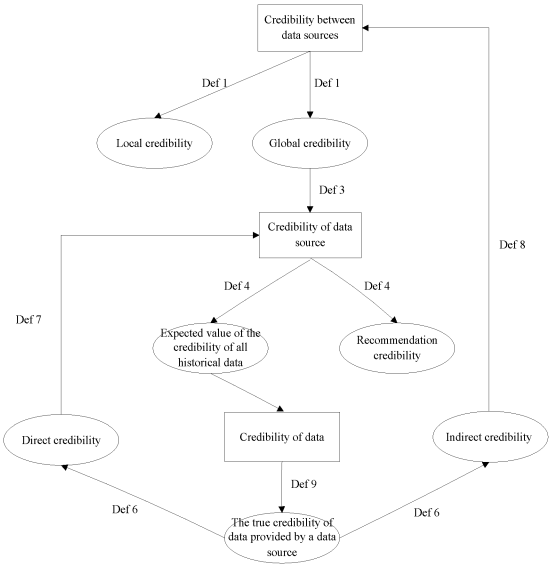

Fig. 1. The correlation of the credibility definition between data sources, data, data source

IV. CASE ANALYSIS AND VERIFICATION

In order to validate the practicability of the proposed method in this paper, the object of the simulation experiment is the data of the real time operation of a power company's dispatching and communication center, which includes the local power grid data of three power plants and four substations. In this system, the total amount of each group is obtained from 108 measurements, including 12 node voltages, 10 active and reactive powers of generation units, 15 load flow values, 3 active and reactive powers of transformer input and 10 line power flow values. Each measure value is assigned to each label to facilitate further analysis.

A. Design of simulation experiment

In this experiment, There are more parameters involved in the model. The collected data is divided into two parts. A part of data is used to establish the power trusted network, which is trained repeatedly and adjusts the value of the parameters, the other part of data is to verify the stability and accuracy of the model.

B. Design of simulation experiment

As mentioned above, firstly, the credibility of an entity relative to other entities is calculated, starting from the formula (1) to calculate the credibility between data sources. According to the formula (2) and formula (3), the contents of the two aspects are calculated, on the one hand, the formula needs to calculate the local credibility. If the entity has a context interaction (condition 1) or a new behavior (condition 2), the local credibility need updating, if there is no new behavior, time penalty is imposed on it. If the entity meets the condition 1 in the calculation process, or the entity not only meets the condition 2, the similarity of data or behaviors provided by between entities but also exceeds a system threshold, the link of the directed graph can be established among entities, thereinto, the weight of the link is the value of local credibility. On the other hand, the formula needs to calculate the global credibility.

Secondly, the credibility of an entity is calculated by the formula (4). If the expected value of the credibility of all historical data provided by the entity or the recommendation credibility of entities of each layer changes, the credibility of the entity is updated. If the credibility is not changed, time penalty is imposed on it.

Finally, the formula (7) calculates the credibility of data depending on a theme by using the probability of complementary events. The formula (8) gives the true credibility of data provided by the entity. Meanwhile, the formula (9) gives the direct credibility of data provided by the entity, and the formula (10) gives the indirect credibility of data provided by the adjacent entities. If an entity provides some malicious and false data in the experiment, the entity will be severely punished, so that it can be a very low value in the trusted network. If its behavior is always normal, the credibility will be improved with the increase of their credit.

C. Experimental results and analysis

Combined with 4.2 section, data are imported into the algorithm to verify the feasibility. In the process of experiment, we artificially set a power equipment's data, which are mainly to verify the model for the detection of false data, processing capacity, using the formula (1), formula (4), formula (7) for calculating the credibility, and observe the change of its credibility with time. As shown in Figure 2.

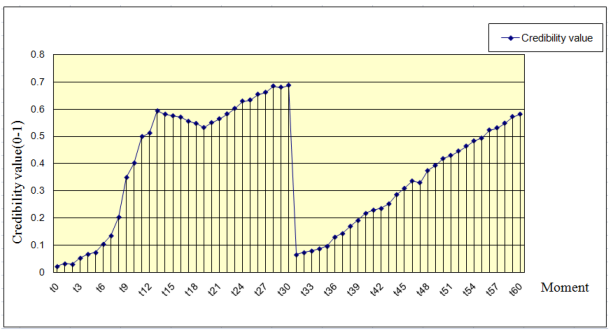

Fig. 2. Artificial power equipment's trusted value changing with time trend

From Figure 3, we can find that the equipment's credibility presents a rising trend at the T_0-T_{30} moment, but the equipment's credibility has a slow downward trend at the T_{12}-T_{18} moment, which is mainly due to the absence of new behavior and imposes time penalty on its credibility.

As a result of the equipment to make an unreliable behavior at the T_{31} moment, it has been punished, resulting in its credibility dropped to 0.1. After the T_{32} moment, the equipment's behavior is normal and the credibility starts recovering, but the trend is relatively slow.

The trusted network topology diagram of hierarchical data sources and the transitive diagram of the credibility of hierarchical data sources relative to a data are shown in Figure 5 at a certain moment.

By definition 2, the formula (2) is used to calculate the local credibility between data sources in smart grid, and the trusted network can be constructed. As shown in Figure 3 (a), a partial network topology graph is given, as shown in Figure 3 (b), the transitive credibility diagram is given for certain data. We can draw from the fact that a data not only has direct contact with the provider, but also is surrounded by a lot of data sources which are directly or indirectly linked to the data, forming a small trusted network, which can greatly improve the accuracy of a data credibility evaluation.

V. CONCLUSIONS

In this paper, the typical characteristics and attributes of smart grid big data are analyzed in detail with combination of between the bad data identification models and the credibility analysis' models of general data. Based on the hierarchical model, it gives the analysis model of smart grid big data's credibility measurement. In the case of the large amount of data provided by data sources, the model can accurately analyze the credibility of data, and it is better to satisfy the requirement of Big Data. A simple instance is selected to verify the feasibility of the model. But the method of building the credibility analysis' network still needs to be improved. Above three points will be the focus in further research work.

References

[1] Hao Jinping, Piechocki Robert J, Kaleshi Dritan, Chin Woon Hau, Fan Zhong. "Sparse Malicious False Data Injection Attacks and Defense Mechanisms in Smart Grids", IEEE Transactions on Industrial Informatics, vol. 11, no. 05, pp. 1198-1209, 2015.

[2] K. L. GAO, Y. ZH. XIN, ZH. LI, et al. "Development and Process of Cybersecurity Protection Architecture for Smart Grid Dispatching and Control Systems", Automation of Electric Power Systems, vol. 39, no. 01, pp. 48-52, 2015.

[3] J. L. Duran-Paz, F .Perez-Hidalgo, M. J. Duran-Martinez. "Bad Data Detection of Unequal Magnitudes in State Estimation of Power Systems", IEEE Power Engineering Review, vol. 121, no. 05, pp. 57-60, 2002.

[4] Shyh-Jier Huang, Jeu-Min Lin. "Enhancement of Anomalous Data Mining in Power System Predicting-Aided State Estimation", IEEE Transaction on Power System, vol. 19, no. 01, pp. 610-619, 2004.

[5] SH. R. WANG, L. ZHANG and H. S. LI. "Evaluation Approach of Subjective Trust Based on Cloud Model", Journal of Software, vol. 21, no. 06, pp. 1341-1352, 2010.

[6] W. TANG and ZH. CHEN. "Research of Subjective Trust Management Model Based on the Fuzzy Set Theory", Journal of Software, vol. 14, no. 08, pp. 1401-1408, 2003.

[7] T. ZHANG, M. Y. ZHAI, H. B. ZHANG, et al. Substation Topology Error Identification Based on Uncer-tainty Reasoning, Automation of Electric Power Systems, vol. 38, no. 06, pp. 49-54, 2014.

[8] L. ZHANG, J. W. LIU, R. CH. WANG and H. Y. WANG. "Trust evaluation model based on improved D-S evidence theory", Journal on Communications, vol. 34, no. 07, pp. 167-173, 2013.

[9] Q. Y. ZHAO, W. L. ZUO, ZH. SH. TIAN and Y. WANG. "A Method for Assessment of Trust Relationship Strength Based on the Improved D-S Evidence Theory", Chinese Journal of Computers, vol. 37, no. 04, pp. 873-883, 2014.

[10] J. ZHOU, Q. WANG, C. C. HUNG and X. J. Yi. "Credibilistic Clustering: The Model and Algorithms", INTERNATIONAL JOURNAL OF UNCERTAINTY FUZZINESS AND KNOWLEDGE-BASED SYSTEMS, vol. 23, no. 04, pp. 545-564, 2015.

[11] He Kemeng, Sun Yushan,Qin Zhang, Sun Yuqiang and Gu Yuwan. "The research of trusted attribute based on model-driven of MDA", Open Electrical and Electronic Engineering Journal, vol. 08, no. 01, pp. 273-277, 2014.

[12] W. Yu, S. J. Li, S. Yang, Y. H. Hu, J. Liu, Y. G. Ding and H. Du, "Automatically Discovering of Inconsistency Among Cross-Source Data Based on Web Big Data", Journal of Computer Research and Development, vol. 52, no. 02, pp. 295-308, 2015.

[13] Hsiao-Ling Wu and Chin-Chen Chang, "A Robust Image Encryption Scheme Based on RSA and Secret Sharing for Cloud Storage Systems", Journal of Information Hiding and Multimedia Signal Processing, vol. 6, no. 02, pp. 288-296, 2015.

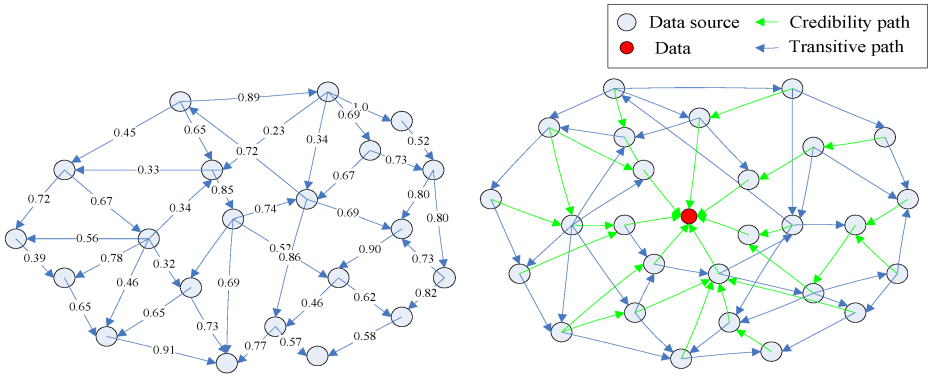

Fig. 3(a) A partial network topology graph

Fig. 3(b) Transitive credibility diagram for a certain data

Fig. 3. The partial topological diagram of trusted network at a certain time