USI-AUC : an evaluation criterion of community detection based on a novel link-prediction method

Citation for published version (APA):

Wu, Y., Yu, H., Zhang, J., Liu, S., Huang, R., & Li, P. (2018). USI-AUC : an evaluation criterion of community detection based on a novel link-prediction method. *Intelligent Data Analysis*, 22(2), 439-462.
<https://doi.org/10.3233/IDA-173400>

DOI:

[10.3233/IDA-173400](https://doi.org/10.3233/IDA-173400)

Document status and date:

Published: 14/03/2018

Document Version:

Accepted manuscript including changes made at the peer-review stage

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy

If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/321494900>

USI-AUC: an evaluation criterion of community detection based on a novel link-prediction method

Article in *Intelligent Data Analysis* · December 2017

CITATIONS

0

READS

14

6 authors, including:

[Jianpeng Zhang](#)

Technische Universiteit Eindhoven

12 PUBLICATIONS 13 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:

Sampling Clustering Process [View project](#)

Community Detection [View project](#)

USI-AUC: an evaluation criterion of community detection based on a novel link-prediction method

Yiteng Wu^a, Hongtao Yu^{a,*}, Jianpeng Zhang^{a,b}, Shuxin Liu^a, Riyang Huang^a and Peng Li^a

^a *National Digital Switching System Engineering and Technological R&D Center*

Zhengzhou, 450002, P. R. China

^b *Eindhoven University of Technology*

5600 MB Eindhoven, the Netherlands

**hunter-yxq@126.com*

Modularity Evaluation (ME) is usually used in community detection for evaluating the disjoint and overlapping communities. In this paper, two obvious defects of ME are revealed and proved, including the non-decreasing contribution of isolated nodes to modularity and lacking of appropriate measures on overlapping community. We also propose a new evaluation criterion, the USI-AUC, which is the Area Under the Curve (AUC), originated from link-prediction of Uniform-Structure-Information (USI) model. We test the new criterion on various datasets, and find that such criterion can avoid the issues exposed in ME.

Keywords: evaluation of communities; modularity evaluation; USI model; link prediction; USI-AUC criterion.

1. Introduction

Community structure is an important property in complex networks, with a variety of different types of networks showing strong community effects in society and nature,^{1,2} especially in this Big Data era.³ Taking the social software WeChat as an example, various individuals with common interests form multiple WeChat groups.

In recent years, researches in community detection have been very active, and how to discover valuable community structures in a network remains a hot topic in the field.⁴ Researchers have proposed numerous community detection algorithms, which are mainly divided into two categories according to whether they allow overlapping, i.e., disjoint-community detection and overlapping-community detection. For disjoint-community detection, modularity optimization,^{5,6} spectral-clustering,^{7,8} hierarchical-clustering,⁹ and label-propagation methods are proposed.^{10,11} The disjoint-community detection algorithm was extended by Palla et al.,¹² which allows nodes to belong to different communities. Thereby, it expands the field of overlapping-community researches. Typical methods include the clique-percolation,¹³ line graph and link partition,¹⁴ local expansion and optimization,¹⁵ and fuzzy detection methods.¹⁶

*Corresponding author.

With the continuous development of community detection algorithms, evaluation criterion of communities has also become a research hotspot. Because of the unknown community information in the actual network, evaluation strategies from various angles are utilized. However, there is no generally accepted evaluation method for community detection. There exist different evaluation criteria, such as Newman’s typical-modularity criterion⁵ and omega index for disjoint communities.^{17,18} For the evaluation of overlapping communities, Clara et al.¹⁹ proposed an overlapping-community evaluation formula based on vertex and edge density; however, the value increases along with the increasing number of communities. Additionally, some studies extended the evaluation metrics from disjoint communities to overlapping communities. Nicosia et al.²⁰ extended the modularity criterion in directed networks, and Nepusz et al.¹⁶ and Huawei Shen et al.^{21,22} extended the modularity in undirected networks with widespread applications. Modularity and extended modularity are of great significance to study community structure. However, for evaluating communities, the ME (containing modularity and extended modularity criteria) also exhibits theoretical defects, including the increasing contribution of isolated nodes to modularity and lacking of appropriate measures on overlapping community. Taking the overlapping community as an example, Fig. 1 shows the original division and overlapping division of the Zachary dataset.²³ Obviously, it is more reasonable to consider nodes 3,9 as common nodes; however, a kind of extended modularity, EQ , is 0.351, which is lower than the modularity of original division (0.371).

Fig. 1. Diagram of original and overlapping communities. Nodes 3,9 are active in both the left community and the right community; therefore, it is reasonable to serve nodes 3,9 as common nodes in both communities. However, EQ cannot provide a reasonable evaluation.

Though this phenomenon was mentioned in Ref. 24, they did not do any theoretical analysis. In this study, the theoretical defects of ME are analyzed in detail through two theorems accompanied by theoretical proofs and data verification. In the analysis of the theoretical defects of ME, the area under the curve (AUC) accuracy of the link-prediction method based on a Uniform-Structure-Information (USI) model is proposed as an evaluation criterion defined as USI-AUC. This criterion allows a unified and effective evaluation of the disjoint and overlapping communities and we find that such criterion can avoid the issues exposed in ME.

The remainder of the paper is organized as follows. In Section 2, we report the related work on evaluating overlapping communities, especially the evaluation criteria without known real community, i.e., ME. In Section 3, we propose two theorems to verify the defects of ME. In Section 4, we present the USI-AUC-evaluation criterion based on the link-prediction method. In Section 5, we provide experimental verification of the two theorems and comparative analysis of USI-AUC and other criteria.

2. Related work

Unlike disjoint community detection, where a number of measures have been proposed for comparing identified partitions with the known partitions, only a few measures are suitable for a set of overlapping communities.²⁴ These measures are mainly divided into 2 categories, one is based on the data of labeled network, the other is the evaluation criterion which does not need the labeled data.

2.1. Evaluation criterion based on labeled network

In the case of labeled network, accuracy evaluation criterion of precision, recall and F1-score can be used to evaluate the community detection algorithm. Precision is the number of correctly detected overlapping nodes N_c divided by the total number of detected overlapping nodes N_a ; Recall is the number of correctly detected overlapping nodes N_c divided by the true number of overlapping nodes N ; F1-score is the harmonic mean of precision and recall.

$$\text{precision} = \frac{N_c}{N_a}, \quad (1)$$

$$\text{recall} = \frac{N_c}{N}, \quad (2)$$

$$\text{F1-score} = \frac{2\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \quad (3)$$

Lancichinetti¹⁵ has extended the notion of Normalized Mutual Information (NMI)²⁵ within the framework of information theory to account for the evaluation of overlapping communities. Omega Index¹⁷ is the overlapping version of the Adjusted Rand Index. It is based on pairs of nodes in agreement in two communities. For labeled network datasets, NMI and omega are most widely used measures.

2.2. Evaluation criterion without labeled network

In community detection studies, most occasions we do not know the true communities, so the evaluation criteria based on modularity are usually adopted. In the

unweighted-undirected network, the value of modularity is defined as the proportion of inner edges in communities in the network minus an expected value which is the proportion of inner edges in communities under the same community distribution when the network is set to random network.⁵ The underlying assumption of modularity metric is that the higher the densities of edges in communities are, the better the result of community division is. Assuming that l communities have been divided as $P = \{C_1, C_2, \dots, C_l\}$, the modularity Q is defined as

$$Q = \frac{1}{2m} \sum_{i=1}^l \sum_{v \in C_i, w \in C_i} \left(A_{vw} - \frac{k_v k_w}{2m} \right), \quad (4)$$

where A_{vw} is the adjacency matrix. m represents the numbers of edges of network. v, w means the number of node. k_v expresses the degree of node v , and C_i expresses the article i community.

Fig. 2. Example of modularity calculation.

As shown in Fig. 2, the network contains 8 nodes and 11 edges, and its adjacency matrix is

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}, \quad (5)$$

The node degree of v_1 is $k_1 = \sum_{w=1}^8 A_{1w} = 3$. Similarly, $k_2 = 2, k_3 = 2, k_4 = 5, k_5 = 4, k_6 = 2, k_7 = 2, k_8 = 2$. If the network is divided into two communities,

according to Eq. 4 the modularity Q is calculated as:

$$\begin{aligned}
 Q = \frac{1}{2 \times 11} \left\{ \left[\left(3 - \frac{3 \times (3+2+2+5)}{22} \right) + \left(2 - \frac{2 \times (3+2+2+5)}{22} \right) \right. \right. \\
 + \left. \left(2 - \frac{2 \times (3+2+2+5)}{22} \right) + \left(3 - \frac{5 \times (3+2+2+5)}{22} \right) \right] \\
 + \left[\left(3 - \frac{4 \times (4+2+2+2)}{22} \right) + \left(1 - \frac{2 \times (4+2+2+2)}{22} \right) \right. \\
 \left. \left. + \left(2 - \frac{2 \times (4+2+2+2)}{22} \right) + \left(2 - \frac{2 \times (4+2+2+2)}{22} \right) \right] \right\} = 0.3140
 \end{aligned} \quad (6)$$

For overlapping-community detection, Nepusz et al.¹⁶ proposed an extension method, Q_f , to the modularity, Q , by weighting each node associating with a community:

$$Q_f = \frac{1}{2m} \sum_{i=1}^l \sum_{v \in C_i, w \in C_i} \alpha_{iv} \alpha_{iw} \left(A_{vw} - \frac{k_v k_w}{2m} \right), \quad (7)$$

where

$$\alpha_{iv} \in [0, 1] \text{ for all } 1 \leq i \leq l, 1 \leq v \leq N,$$

$$\sum_{i=1}^l \alpha_{iv} = 1 \text{ for all } 1 \leq v \leq N,$$

$$0 < \sum_{v=1}^N \alpha_{iv} < N \text{ for all } 1 \leq i \leq l.$$

The total number of nodes in the network is N , and the weighted coefficient, α_{iv} , is a measure of the strength of association between node v and community i . When the weighted coefficients are all 1, the extended modularity, Q_f , degenerates into the modularity, Q .

To simplify the computing process of weighted coefficient, Huawei Shen et al.²¹ proposed a specific extension of the modularity, EQ :

$$EQ = \frac{1}{2m} \sum_{i=1}^l \sum_{v \in C_i, w \in C_i} \frac{1}{O_v O_w} \left(A_{vw} - \frac{k_v k_w}{2m} \right), \quad (8)$$

where O_v expresses the number of communities in which node v is subordinate.

3. The theoretical defects of ME

Based on the definition of modularity, Q , and extended modularity, Q_f or EQ , we propose the following definition:

Definition 1. The contribution of a single node, v_0 , to the extended modularity, Q_f , in an arbitrary community, C_i , is defined as

$$\Delta_{v_0}^{C_i} = \frac{1}{2m} \sum_{w \in C_i} \alpha_{iv_0} \alpha_{iw} \left(A_{v_0w} - \frac{k_{v_0} k_w}{2m} \right) I_{C_i}(v_0), \quad (9)$$

where

$$I_{C_i}(x) = \begin{cases} 1, & x \in C_i \\ 0, & x \notin C_i \end{cases}. \quad (10)$$

It is easy to know from the definition 1, $Q_f = \sum_i \sum_j \Delta_{v_j}^{C_i}$.

Fig. 3. Example of calculating the contribution of a single node to the extended modularity.

Take Fig. 3 as an example, the modularity of the contribution of node v_4 to the community C_1 is

$$\begin{aligned} \Delta_{v_4}^{C_1} &= \frac{1}{2 \times 11} \times \left[\alpha_{14} \alpha_{11} \left(A_{41} - \frac{k_4 k_1}{22} \right) + \alpha_{14} \alpha_{12} \left(A_{42} - \frac{k_4 k_2}{22} \right) \right. \\ &\quad \left. + \alpha_{14} \alpha_{13} \left(A_{43} - \frac{k_4 k_3}{22} \right) + \alpha_{14} \alpha_{14} \left(A_{44} - \frac{k_4 k_4}{22} \right) \right] \\ &= \frac{1}{22} \left(\frac{7}{22} \alpha_{14} \alpha_{11} + \frac{12}{22} \alpha_{14} \alpha_{12} + \frac{12}{22} \alpha_{14} \alpha_{13} - \frac{25}{22} \alpha_{14} \alpha_{14} \right) \end{aligned} \quad (11)$$

If the weight coefficients are given in accordance with EQ, i.e., $\alpha_{14} = \frac{1}{2}$, $\alpha_{11} = \alpha_{12} = \alpha_{13} = 1$, then $\Delta_{v_4}^{C_1} = 0.0191$, similarly, the modularity of the contribution of node v_4 to the community C_2 is $\Delta_{v_4}^{C_2} = -0.0191$ (this value is independent with $\Delta_{v_4}^{C_1} = 0.0191$), and then the contribution of node v_1 to the modularity in community C_2 is 0. The reason is that node v_1 does not belong to community C_2 , $I_{C_2}(1) = 0$.

On the basis of **definition 1** as well as the modularity definition of Newman, and the extension-modularity definition of Nepusz et al.¹⁶ or Huawei Shen et al.,²¹ we obtain the following two theorems.

Theorem 1. *There must be a disjoint community structure, in which an arbitrary single node is a non-negative contributor to the whole modularity, and modularity does not increase when deleting arbitrary nodes in this community structure.*

Specifically, when the node is isolated, deleting the isolated node does not increase modularity.

Proof. Assuming that l communities have been divided as $P = \{C_1, C_2, \dots, C_l\}$, for arbitrary node v_0 , its degree is k_{v_0} . Assume that node v_0 belongs to the article i community. The contribution of v_0 to the whole modularity in community C_i is

$$\Delta_{v_0}^{C_i} = \frac{1}{2m} \sum_{w \in C_i} \left(A_{v_0 w} - \frac{k_{v_0} k_w}{2m} \right), \quad (12)$$

given that

$$\sum_{i=1}^l \sum_{w \in C_i} \left(A_{v_0 w} - \frac{k_{v_0} k_w}{2m} \right) = 0. \quad (13)$$

Therefore, there must be a disjoint community structure, where node v_0 is a non-negative contributor to the whole modularity.

If $\sum_{w \in C_i} \left(A_{v_0 w} - \frac{k_{v_0} k_w}{2m} \right)$ is 0 for all $i = 1, 2, \dots, l$, then the modularity is constant when deleting node v_0 ; otherwise there must be $\sum_{w \in C_i} \left(A_{v_0 w} - \frac{k_{v_0} k_w}{2m} \right) > 0$, resulting in decreased modularity when deleting node v_0 . In summary, modularity does not increase when deleting arbitrary nodes in this community structure. $\square$

This theorem indicates that modularity is sensitive to isolated nodes, although the division of isolated nodes is reasonable, the criterion of modularity also decreases, which cannot reflect the rationality in the division of isolated nodes.

Lemma 1. *If moving a node to a different community does not cause changes in the weighted coefficients of other nodes, the influence of the node on the whole modularity is only related to the change in contribution of this node, and the subsequent change in contribution of the other nodes remains the same as that of this node.*

Proof. Upon changing the community to which arbitrary node v_0 belongs, the change in contribution to modularity of arbitrary community C_i is

$$\begin{aligned}
& \sum_{w \in C_i^{new}} \Delta_w^{C_i^{new}} - \sum_{w \in C_i^{old}} \Delta_w^{C_i^{old}} \\
&= \Delta_{v_0}^{C_i^{new}} - \Delta_{v_0}^{C_i^{old}} + \sum_{w \in C_i^{new} \setminus v_0} \Delta_w^{C_i^{new}} - \sum_{w \in C_i^{old} \setminus v_0} \Delta_w^{C_i^{old}} \\
&= \frac{1}{2m} \sum_{w \in C_i^{new} \setminus v_0} \alpha_{iv_0}^{new} \alpha_{iw}^{new} \left(A_{v_0 w} - \frac{k_{v_0} k_w}{2m} \right) I_{C_i^{new}}(v_0) \\
&\quad - \frac{1}{2m} \sum_{w \in C_i^{old} \setminus v_0} \alpha_{iv_0}^{old} \alpha_{iw}^{old} \left(A_{v_0 w} - \frac{k_{v_0} k_w}{2m} \right) I_{C_i^{old}}(v_0) \\
&\quad + \frac{1}{2m} \sum_{v \in C_i^{new} \setminus v_0} \sum_{w \in C_i^{new} \setminus v} \alpha_{iv}^{new} \alpha_{iw}^{new} \left(A_{vw} - \frac{k_v k_w}{2m} \right) I_{C_i^{new}}(v) \\
&\quad - \frac{1}{2m} \sum_{v \in C_i^{old} \setminus v_0} \sum_{w \in C_i^{old} \setminus v} \alpha_{iv}^{old} \alpha_{iw}^{old} \left(A_{vw} - \frac{k_v k_w}{2m} \right) I_{C_i^{old}}(v) \\
&= 2 \left(\Delta_{v_0}^{C_i^{new}} - \Delta_{v_0}^{C_i^{old}} \right), \tag{14}
\end{aligned}$$

where $\Delta_w^{C_i^{new}}$ expresses the contribution of node w to Q_f in community C_i after moving node v_0 and $\Delta_w^{C_i^{old}}$ expresses the contribution of node w to the Q_f in community C_i before moving node v_0 . Derivation results showed that changes in modularity of community C_i were only related to changes in the contribution of node v_0 .

Because Eq. (14) holds for arbitrary community C_i , the influence of node v_0 on the other involved communities is only related to the change in contribution of node v_0 to the involved communities; the influence of node v_0 on the whole modularity is only related to the change in contribution of node v_0 . $\square$

Lemma 1 shows that investigating the impact of moving a node to a different community in the absence of changes in the coefficients of other nodes is only necessary to consider changes in the contribution of the coefficient of the moved node without regard for the subsequent change in contribution of other nodes.

For example, in Fig. 3, moving the common node v_4 to community C_1 and C_2 , respectively, according to **definition 1**, not only the contribution of node v_4 to the modularity is changed, the contributions of nodes $v_1, v_2, v_3, v_5, v_6, v_7, v_8$ in the communities where node v_4 moves before and after are changed. The significance of this lemma is that we only need to examine the node v_4 on their own contribution to the change of modularity, without regard to the contributions of other nodes, and facilitate the proof of **Theorem 2**. In this example, the influence of node v_4 on the whole modularity when moving the common node v_4 to community C_1 is

$2 \left(\Delta_{v_4}^{C_1^{new}} - \Delta_{v_4}^{C_1^{old}} \right) + 2 \left(\Delta_{v_4}^{C_2^{new}} - \Delta_{v_4}^{C_2^{old}} \right) = 2(0.0382 - 0.0191) + 2(0 - (-0.0191)) = 0.0764$. The influence of node v_4 on the whole modularity when moving the common node v_4 to community C_2 is $2 \left(\Delta_{v_4}^{C_1^{new}} - \Delta_{v_4}^{C_1^{old}} \right) + 2 \left(\Delta_{v_4}^{C_2^{new}} - \Delta_{v_4}^{C_2^{old}} \right) = 2(0 - 0.0191) + 2(-0.0382 - (-0.0191)) = -0.0764$.

Theorem 2. *There must be a disjoint community in which modularity is no less than that of any given overlapping community.*

Proof. For any given overlapping community, assuming that arbitrary nodes $v_1, v_2, \dots, v_p$ serve as common nodes in overlapping communities, each node belongs to $q_{v_1}, q_{v_2}, \dots, q_{v_p}$ overlapping communities, respectively. Then, the contribution of common node v_j to the whole modularity is

$$\Delta_{v_j} = \frac{1}{2m} \sum_{i=1}^{q_{v_j}} \sum_{w \in C_{v_j}^i} \alpha_{iv_j} \alpha_{iw} \left(A_{v_j w} - \frac{k_{v_j} k_w}{2m} \right) \quad (15)$$

where $C_{v_j}^i$ expresses the article i community containing node v_j . The contribution of node v_j to the whole modularity in community $C_{v_j}^i$ is

$$\Delta_{v_j}^{C_{v_j}^i} = \frac{1}{2m} \sum_{w \in C_{v_j}^i} \alpha_{iv_j} \alpha_{iw} \left(A_{v_j w} - \frac{k_{v_j} k_w}{2m} \right) \quad (16)$$

Now, inspect the common nodes one by one.

- (i) When moving one node (e.g., node v_1) does not cause changes in the weighted coefficients of other nodes, according to **lemma 1**, we should only consider the change of contribution of node v_1 . In the q_{v_1} overlapping communities to which node v_1 belongs, there must be a community in which node v_1 has the maximum contribution to modularity. Assuming the contribution in the first overlapping community $C_{v_1}^1$ reaches the maximum, and the contribution is $\Delta_{v_1}^{C_{v_1}^1}$, according to Eq. (16), moving node v_1 as a non-common node to community $C_{v_1}^1$ does not reduce modularity. Therefore, after moving node v_1 , the modularity of the overlapping community composed of $(p-1)$ common nodes $\geq$ the modularity of the overlapping community composed of p common nodes, and the whole modularity is not reduced. Similarly, after moving common nodes as non-common nodes $(p-1)$ times, the overlapping community is converted to a disjoint community without reducing the whole modularity.
- (ii) When moving one node (e.g., node v_1) causes changes in the weighted coefficients of other nodes, inspect the common nodes separate from the whole. The influence on modularity of moving these common nodes together is equivalent to the influence of moving these common nodes individually, assuming that moving one node does not cause changes in the

weighted coefficients of other nodes. From (i), we know that the modularity is not reduced when moving these nodes as non-common nodes and thereby creating a disjoint community. $\square$

Theorem 2 shows that there is no reasonable overlapping community from the perspective of modularity. If we once had a reasonable overlapping community, it would be possible to obtain a more reasonable disjoint community in the view of modularity by adjusting common nodes. Therefore, even if the division of the overlapping community is reasonable, it is difficult to give a rational evaluation using Q_f and EQ .

According to this theorem, moving the common node v_4 to community C_1 and C_2 , respectively, there must be a way to move the node to get a greater modularity. The community is shown as Fig. 4 after moving node v_4 to community C_1 and C_2 . If the weight coefficients are given in accordance with EQ, the modularity which node v_4 serves as common node is 0.2376, and modularity changes to $0.2376 + 0.0764 = 0.3140$ and $0.2376 - 0.0764 = 0.1612$ when moving it to community C_1 and C_2 , respectively.

Fig. 4. An example to explain the lemma.

To address the theoretical defects of modularity evaluation, we present an evaluation criterion using the accuracy of a novel link prediction method.

4. USI-AUC-evaluation criterion for community detection

There is a strong positive correlation between the community division predicted by the community detection algorithm and link prediction. If link prediction depends only on community structure information, the more reasonable the community structure, the more accurate link prediction will be. Conversely, the more precise the link prediction, the more rational the community structure will be. Based on this analysis, link prediction can be performed using only community information, taking the AUC determined by link prediction as the evaluation criterion for community detection.

The existing methods for link prediction using community information are divided into the following two types: calculating the similarity index from the whole network to the specific community interior^{26,27} and regarding the community information as a kind of property information and weighting it to the similarity index.²⁸ Both of these methods regard community information as supplementary information in order to improve the accuracy of link prediction. However, the link-prediction method for evaluating the performance of community detection should take only the community information into consideration without other information. Therefore, the above two methods are not applicable to the evaluation of community detection. This paper presents an evaluation criterion based on link-prediction using the USI model that adopts only community information and uses AUC accuracy to evaluate community detection (USI-AUC). This method can be applied to assess disjoint, overlapping, and hierarchical communities among others.

First, we provide the definition of the USI model, then present to the link prediction algorithm based on the USI model, followed by evaluation of the USI-AUC metric. This process is illustrated in Fig. 5.

Fig. 5. Diagram of the USI-AUC-evaluation criterion for community detection.

4.1. The definition of Uniform-Structure-Information model

Definition 2. USI model.

- (i) A_0 represents the set of all vertexes in the network. Define the power set as:

$$A_1 = 2^{A_0}, A_2 = 2^{A_1}, A_3 = 2^{A_2}, \dots, A_i = 2^{A_{i-1}}, i = 1, 2, 3, \dots$$

For arbitrary i , elements with specific relationship, f , consist of a family set $\mathfrak{D}$, $\mathfrak{D} = \{D_1, D_2, \dots, D_n\}$:

$$f : (A_i)^k \rightarrow \mathfrak{D}, \mathfrak{D} \subset (A_i)^k, k = 1, 2, \dots, |A_i|, \quad (17)$$

where

$$(A_i)^k = \{\{A_{i1}, A_{i2}, \dots, A_{ik}\} : A_{ij} \in A_i, 1 \leq j \leq k\}.$$

Specifically, when $i=0$,

$$f : (A_0)^k \rightarrow \mathfrak{D}_0, \mathfrak{D}_0 \subset (A_0)^k.$$

(ii) Define a discrete metric space based on the elements of family set $\mathfrak{D}$:

$$d_i(x, y) = \begin{cases} p_i(p_i \leq 1), x \neq y \text{ and } x, y \in D_i \\ 0, x = y \text{ and } x, y \in D_i \end{cases}. \quad (18)$$

The USI model establishes a concise description for hierarchical and overlapping relationships of nodes in the network and can explain various relationships between the connection and organization of nodes in the network. **Definition 2** (i) describes that nodes or sets can consist of a new set in which elements have specific relationships or properties. In **Definition 2** (ii), the metric p in the discrete metric space can be interpreted as the probability that internal elements come in contact with each other, with this probability identical for all elements within a given set.

From the definition of the USI model, the elements of a set could consist of sets or non-sets. When the elements are non-sets, the metric p expresses the probability between nodes directly. When the elements are sets, metric p expresses the probability between sets. For example, all the students in a class comprise a set consisting of non-sets. The metric p describes the contact probability for any two students in the class, as well as for all remaining students. As another example, all the classes in a school comprise another set consisting of sets. In this case, the metric p describes probabilities between any two classes, thus indirectly implicating contact probabilities between any two arbitrary students between classes.

Definition 3. The order of elements.

In the USI model, elements of A_k and its nonempty subsets are called k -order elements ($k = 0, 1, 2, 3, \dots$). Elements where $k \geq 2$ order are called high-order elements.

Definition 4. The order of sets.

In the USI model, the order of elements contained within a set is referred to as set order.

For example, considering a network with 3 nodes, $A_0 = \{1, 2, 3\}$. Obviously, the elements 1, 2, 3 are all 0-order elements according to definition 3, and A_0 is a 0-order set based on definition 4. Because the element $\{1, 2\} \in 2^{A_0} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\} = A_1$, element $\{1, 2\}$ is a 1-order element. Naturally, $\{\{1, 2\}\}$ is a 1-order set, owing to the 1-order element $\{1, 2\}$ as the basic of definition 4. As another example, $\Lambda_1 = \{\{1, 2\}, \{1, 3\}, \{1, 2, 3\}\}$ is a nonempty subset of A_1 , $\Lambda_1 \subseteq A_1$, so the element of Λ_1 , e.g., $\{1, 2, 3\}$, is a 1-order element, and Λ_1 is a 1-order set. Similarly, $\{\{1, 3\}, \{1, 2, 3\}\} \in 2^{\Lambda_1} \subseteq 2^{A_1} = A_2$, therefore the element $\{\{1, 3\}, \{1, 2, 3\}\}$ is a 2-order element.

The USI model is a generalization of the hierarchical-structure model²⁹ and the stochastic block model.^{30, 31} From the USI model, if and only if each set contains two elements and each element belongs only to the corresponding order set, the USI model degenerates to a hierarchical-structure model that could contain high-order

elements. When the highest order of a set is one and the intersection of all elements is the empty set, the USI model degenerates to a stochastic block model.

When the specific relationship describes that the division results of the community detection algorithm describe elements consisting of a set, the AUC of link prediction for the USI model is used as the evaluation criterion for the community detection algorithm. Because the USI model does not require that 0-order elements belong to the set, this criterion can evaluate both the disjoint and overlapping communities.

Proposition 1. A $k(k \geq 1)$ -order set $\mathfrak{X}^k = \{X_1^k, X_2^k, \dots, X_n^k\}$ can be reduced in order to $(k - 1)$ by performing union operations on elements:

$$g : \mathfrak{X}^k \rightarrow \bigcup_{i=1}^n X_i^k. \quad (19)$$

Clearly, $\bigcup_{i=1}^n X_i^k$ is a $(k - 1)$ -order set and marked as $\mathfrak{X}^{k-1}$.

Corollary 1. A $k(k \geq 2)$ -order set $\mathfrak{X}^k = \{X_1^k, X_2^k, \dots, X_n^k\}$ can be reduced in order to a 0-order set by performing k -iterations of the element-union operation g .

Proof. According to **proposition 1**, a k -order set can be reduced in order to a $(k - 1)$ -order set by performing the element-union operation g . The $(k - 1)$ -order set can be reduced in order to a $(k - 2)$ -order set by subsequent operations and so on down to a 0-order set. $\square$

For example, we assume a 3-order set,

$$\begin{aligned} \mathfrak{X}^3 &= \{X_1^3, X_2^3, X_3^3\} \\ &= \{\{\{1, 2, 3\}, \{6\}\}, \{\{4, 5\}\}, \{\{\{6\}, \{7, 8, 9\}\}, \{\{\{6, 9\}, \{9\}\}\}\}. \end{aligned} \quad (20)$$

Evidently, $X_1^3 = \{\{1, 2, 3\}, \{6\}\}$, $X_2^3 = \{\{4, 5\}\}$, $X_3^3 = \{\{\{6\}, \{7, 8, 9\}\}, \{\{\{6, 9\}, \{9\}\}\}$ are 3-order elements. According to **corollary 1**, the 3-order set is reduced to 2-order set,

$$\begin{aligned} \mathfrak{X}^2 &= \bigcup_{i=1}^3 X_i^3 = \{\{\{1, 2, 3\}, \{6\}\}, \{\{4, 5\}\}\} \cup \{\{\{6\}, \{7, 8, 9\}\}\} \cup \{\{\{6, 9\}, \{9\}\}\} \\ &= \{\{\{1, 2, 3\}, \{6\}\}, \{\{4, 5\}\}, \{\{6\}, \{7, 8, 9\}\}, \{\{6, 9\}, \{9\}\}\}, \end{aligned} \quad (21)$$

the 2-order set is reduced to 1-order set,

$$\begin{aligned} \mathfrak{X}^1 &= \bigcup_{i=1}^4 X_i^2 = \{\{\{1, 2, 3\}, \{6\}\} \cup \{\{4, 5\}\} \cup \{\{6\}, \{7, 8, 9\}\} \cup \{\{6, 9\}, \{9\}\}\} \\ &= \{\{1, 2, 3\}, \{6\}, \{4, 5\}, \{7, 8, 9\}, \{6, 9\}, \{9\}\}, \end{aligned} \quad (22)$$

and the 1-order set is reduced to 0-order set,

$$\begin{aligned} \mathfrak{X}^0 &= \bigcup_{i=1}^3 X_i^1 = \{1, 2, 3\} \cup \{6\} \cup \{4, 5\} \cup \{7, 8, 9\} \cup \{6, 9\} \cup \{9\} \\ &= \{1, 2, 3, 4, 5, 6, 7, 8, 9\}. \end{aligned} \quad (23)$$

Corollary 2. A $k(k \geq 2)$ -order set $\mathfrak{X}^k = \{X_1^k, X_2^k, \dots, X_n^k\}$ in which elements can represent a $(k-1)$ -order set can be reduced in order 1 by $(k-1)$ iterations of the element-union operation g on $(k-1)$ -order sets X_1 to X_n .

Proof. Regarding the elements of k -order sets as the $(k-1)$ -order set, according to **corollary 1**, each $(k-1)$ -order set can be reduced in order to a 0-order set by $(k-1)$ iterations of the element-union operation g , thereby reducing the original k -order set to a 1-order set consisting of 0-order sets as elements and we can obtain the 1-order set. $\square$

According to **corollary 2**, $\mathfrak{X}^3 = \{X_1^3, X_2^3, X_3^3\}$, The 3-order elements X_1^3, X_2^3, X_3^3 are reduced to 2-order elements,

$$X_1^2 = \{\{1, 2, 3\}, \{6\}\} \cup \{\{4, 5\}\} = \{\{1, 2, 3\}, \{6\}, \{4, 5\}\},$$

$$X_2^2 = \{\{6\}, \{7, 8, 9\}\},$$

$$X_3^2 = \{\{6, 9\}, \{9\}\}.$$

and the 2-order elements X_1^2, X_2^2, X_3^2 are reduced to 1-order elements,

$$X_1^1 = \{1, 2, 3\} \cup \{6\} \cup \{4, 5\} = \{1, 2, 3, 4, 5, 6\},$$

$$X_2^1 = \{6\} \cup \{7, 8, 9\} = \{6, 7, 8, 9\},$$

$$X_3^1 = \{6, 9\} \cup \{9\} = \{6, 9\}.$$

So, we get the 1-order set,

$$\mathfrak{X}^1 = \{X_1^1, X_2^1, X_3^1\} = \{\{1, 2, 3, 4, 5, 6\}, \{6, 7, 8, 9\}, \{6, 9\}\}. \quad (24)$$

4.2. USI-AUC criterion based on a link-prediction method

Here, the process of link prediction was performed on a training set. First, the network set E was divided into two parts, and the training set, E^T , was treated as known information, whereas the testing set (probe set) E^P was used for testing, with no information allowed for prediction. This resulted in $E^T \cup E^P = E$, $E^T \cap E^P = \phi$. In this paper, the training set contained 90% links, with the remaining 10% links consisting of the probe set.

In this section, we regarded the community division of community detection algorithm as the partition of sets in the USI model and estimated the metric p in discrete metric space. We then provided the score for link prediction based on the USI model to obtain the USI-AUC-evaluation criterion.

4.2.1. The partition of sets

For the 0-order sets, $i = 0, k = 1, 2, \dots, |A_0|$, according to the results divided using the community detection algorithm as the specific relationship, f , partition the 0-order sets. Assuming that C_i is a community divided using the community detection algorithm, the division is $P = \{C_1, C_2, \dots, C_n\}$. Therefore, the specific relationship f_1 is

$$f_1 : (A_0)^k \rightarrow \mathfrak{D}_{1k}, \mathfrak{D}_{1k} \subset (A_0)^k, k = 1, 2, \dots, |A_0|, \quad (25)$$

$$\mathfrak{D}_{1k} = f_1[(A_0)^k] = \{C_i : |C_i| = k, C_i \in P\}. \quad (26)$$

For 1-order sets, if the community divided using the algorithm do not contain hierarchical-structure information, considering interactions between two 0-order sets, the specific relationship, f_2 , defines as arbitrary two 0-order sets consisting a 1-order set.

Assume

$$D_i \in \bigcup_{k=1}^{|A_0|} \mathfrak{D}_{1k} = \mathfrak{D}_1. \quad (27)$$

When $i = 1$ and $k = 2$, according to a specific relationship, f_2 :

$$f_2 : (A_1)^2 \rightarrow \mathfrak{D}, \mathfrak{D} \subset (A_1)^2, \quad (28)$$

we obtain

$$\mathfrak{D} = f_2[(A_1)^2] = \{\{D_i, D_j\} : \forall D_i, D_j \in \mathfrak{D}_1, D_i \neq D_j\}. \quad (29)$$

For 1-order and higher-order sets, if the community divided using the algorithm contains hierarchical-structure information, regard the hierarchical information as specific relationships, f_3 , consisting of 1-order or higher-order sets.

$$f_3 : (A_i)^k \rightarrow \mathfrak{D}, \mathfrak{D} \subset (A_i)^k, k = 1, 2, \dots, |A_i|. \quad (30)$$

4.2.2. Estimation of metric p in discrete metric space

According to the difference in set order, the estimation method for metric p is divided into three cases: estimation according to the 0-order set, estimation according to the 1-order set, and estimation according to the high-order set. Specific algorithms are provided as follows.

- (i) The estimation of metric p according to the 0-order set G_0 .

Assume

$$|G_0| = N_1, \quad (31)$$

where $|\cdot|$ is the cardinality of set G_0 , meaning that the number of nodes in set G_0 is N_1 .

In a set consisting of only 0-order elements, the distinction between the elements is whether they are connecting or non-connecting. Metric p describes the connecting probability. Assume that the number of links between nodes in set G_0 is a random variable X , such that X follows a binomial distribution, $B(N, p)$, where N represents the maximum number of possible links, $N = C_{N_1}^2$. Maximum-likelihood estimation is used to estimate metric p . The likelihood function is

$$L(p, x) = p^x (1 - p)^{N - x}, x = 0, 1, 2, \dots, N, \quad (32)$$

where x is the number of actual links between nodes in set A_0 .

Let

$$\frac{dL(p, x)}{dp} = 0. \quad (33)$$

We can obtain

$$\hat{p} = \frac{x}{N}. \quad (34)$$

Fig. 6. Example of estimating the metric p in 0-order set.

As shown in Fig. 6, the 0-order set has 10 nodes and 12 edges, and the estimation of metric p is $\hat{p} = 12/C_{10}^2 = 4/15$.

- (ii) The estimation of metric p according to the 1-order set G_1

$$G_1 = \{S_i : S_i \text{ is 1 order element}, i = 1, 2, 3, \dots\}, \quad (35)$$

$$|G_1| = K. \quad (36)$$

There may be intersections between 1-order elements divided using the overlapping-community detection algorithm:

$$\left| S_i \setminus \bigcap_{i=1}^K S_i \right| = k_i. \quad (37)$$

The maximum number of possible links between 1-order elements is

$$N = \sum_{i \neq j} k_i k_j. \quad (38)$$

Therefore, metric p for set G_1 is defined as the connecting probability between 1-order elements. The problem is also transformed into the estimation of p according to a binomial distribution, $B(N, p)$, with the method the same as that presented in (i).

As Fig. 7(a) shows, there is no intersection of 1-order sets, we only need to consider the actual number of edges (6) and probably the largest number of edges ($6 \times 7 = 42$), so the estimation of metric p is $\hat{p} = 6/(6 \times 7) = 1/7$. And there is an intersection in Fig. 7(b), we only consider the actual number of edges (7) and probably the largest number of edges ($6 \times 7 = 42$) except the intersection. The estimation of metric p is $\hat{p} = 7/(9 \times 5) = 7/45$.

Fig. 7. Example of estimating the metric p in 1-order sets.

- (iii) The estimation of metric p according to the high-order set $G_k (k \geq 2)$

According to **corollary 2**, reduce the order of set G_k to a 1-order set, and then resolve the problem using case (ii).

4.2.3. USI-model-based link-prediction scoring

For the overlapping-community detection algorithm, a node can be subordinate to multiple communities; therefore, there exist multiple channels for two nodes to generate a link, which is very similar to human communications. Each increase in one contacting channel indicates an increase in the probability of two nodes consisting of a link. For the USI model, one node can belong to different sets; therefore, calculate the parallel probability of contacts between node pairs through each channel as a link-prediction score.

$$s_{xy} = 1 - \prod_{i=1}^{N_{xy}} (1 - p_{xy}^i), \quad (39)$$

where s_{xy} represents the final link-prediction score of node pair xy , p_{xy}^i is the contacting probability of node pair xy in an article i common community, and N_{xy} represents the number of common communities in which node pair xy belongs.

4.3. USI-model-based evaluation criterion for community detection

According to the final score of node pair s_{xy} , use the test set to compute the AUC,³² which is the area under the receiver operating characteristic curve, to quantify the accuracy of link prediction. AUC represents the probability that the link-prediction score of a randomly chosen missing link is higher than a nonexistent link. For n independent comparisons, if there are n' comparisons where the missing link returns a higher score and n'' comparisons where the missing and nonexistent links return the same score, define the AUC as

$$AUC = \frac{n' + 0.5n''}{n}. \quad (40)$$

This AUC is defined as the evaluation criterion based on the USI model (USI-AUC). The steps of this algorithm are listed in Table 1.

Table 1. Algorithm steps.

Step1:	Divide the network into training set and test set;
Step2:	According to the results of community detection algorithm, P, follow the Eq. (26), (29) and (30) to partition the 0-order sets, 1-order sets and high-order sets;
Step3:	Estimate the matrix p on each set. 0-order sets and 1-order sets follow the Eq. (34), and high-order sets need to reduce order to 1-order sets according to corollary 2 and follow the Eq. (34) to estimate the matrix p ;
Step4:	Calculate the score s_{xy} of link prediction according to Eq. (39);
Step5:	According to the score of link prediction, compute the AUC of link prediction on the test set, i.e., USI-AUC.

5. Experimental verification

In order to reduce computational complexity and simplify the problem, here, we only considered partitions of 0-order sets. First, the experiment focused on EQ to verify the validity of the theorems. We then comparatively analyzed the evaluation effects of the USI-AUC criterion and the modularity criterion to disjoint and overlapping communities using several datasets. At last we tested the statistical significance for the comparison of the USI-AUC criterion and other criteria.

5.1. Verification of the validity of the two theorems

5.1.1. Experimental verification of Theorem 1

Using the classical Newman-modularity algorithm⁵ on the Dolphin dataset³³ for partitioning communities resulted in five divided communities (see Table 2), with

modularity criterion for this division at $Q = 0.5086$. We deleted one node in the network individually and tested modularity.

Table 2. Partitions of the Dolphin dataset using the Newman-modularity algorithm. Each column indicates a community, and the contents are node numbers.

Community	1	2	3	4	5
	4,9,37,40,60	1,3,8,11,20,29, 31,41,43,45,48	2,6,7,10,14,18, 23,26,27,28,32, 33,42,49,55,57, 58,61	5,12,16,19,22, 24,25,30,36,46, 52,56	13,15,17,21,34, 35,38,39,44,47, 50,51,53,54,59, 62

We observed reductions in modularity after deleting arbitrary nodes in these communities (see Table 3), and verified **theorem 1**.

Table 3. Modularity values after deleting individual nodes in Dolphin dataset, with values representing the modularity after deleting a corresponding node. Each cell corresponds to the node number 1–62 in turn.

Number	1	2	3	4	5	6	7	8	9	10
1–10	0.4899	0.4896	0.4941	0.4973	0.5035	0.4898	0.4803	0.4952	0.4984	0.4756
11–20	0.4889	0.5032	0.5041	0.4708	0.4782	0.4918	0.4812	0.4659	0.4792	0.5004
21–30	0.5049	0.4907	0.5039	0.4997	0.4844	0.4945	0.4945	0.4914	0.5015	0.4814
31–40	0.4826	0.5039	0.4945	0.4751	0.4921	0.5035	0.4987	0.4892	0.4782	0.5032
41–50	0.5045	0.4851	0.4837	0.4828	0.5067	0.4773	0.4995	0.4837	0.5039	0.4995
51–60	0.4954	0.4636	0.5030	0.4995	0.4881	0.4985	0.4992	0.4722	0.5041	0.4917
61–62	0.5039	0.5012								

5.1.2. Experimental verification of Theorem 2

We chose one node as a common node in two communities and tested EQ . We obtained 248 EQ values, with 244 values lower than the original modularity ($Q = 0.5086$), and the other four values exceeding the original modularity (0.5110, 0.5111, 0.5121, and 0.5136, corresponding to nodes 21, 21, 41, and 45, respectively). According to **Theorem 2**, we could obtain a disjoint community exhibiting higher modularity by adjusting common nodes. Moving node 21 from community five to community two, the modularity changed to $Q = 0.5127$, moving node 41 from community two to community five changed the modularity to $Q = 0.5149$, and moving node 45 from community two to community five changed the modularity to $Q = 0.5184$. Moving the three nodes simultaneously changed the modularity to $Q = 0.5191$.

We then chose two nodes as common nodes in two communities and tested the EQ . Using the above adjusting results, we obtained 7560 EQ values, with 7558 values lower than the original modularity ($Q = 0.5191$) and two EQ values exceeding the original modularity. When nodes 21 and 45 served as common nodes,

the EQ achieved a maximum value of $EQ = 0.5216$. According to **Theorem 2**, we transformed the overlapping community to a disjoint community by reversing the two nodes in communities one and five and deleting other common nodes. The modularity became $Q = 0.5232$, which exceeded the maximum value ($EQ = 0.5216$), thereby verifying **Theorem 2**. The adjusted communities are listed in Table 4.

Table 4. Adjusted communities, with each column indicating a community and its node numbers.

Community	1	2	3	4	5
	4,9,37,40,60,21	1,3,8,11,20,29, 31,43,48	2,6,7,10,14,18, 23,26,27,28,32, 33,42,49,55,57, 58,61	5,12,16,19,22, 24,25,30,36,46, 52,56	13,15,17,34,35, 38,39,44,47,50, 51,53,54,59,62, 41,45

It is possible to generalize the case involving two common nodes to cases involving multiple common nodes; therefore, it is unnecessary to enumerate these results here.

5.2. Evaluation of the disjoint-community detection algorithm using USI-AUC criterion

This section comparatively analyzed the evaluation effect of the USI-AUC criterion and modularity criterion for 11 kinds of disjoint-community detection algorithms used on six datasets. The 11 disjoint-community detection algorithms and corresponding parameters are listed as Table 5.

The six datasets are Lesmis,⁴² Zachary,²³ Football,¹ CKM-3,⁴³ PPI-Cell,⁴⁴ and Metabolic.⁴⁵ The basic topological features of 6 real networks are listed in Table 6.

For the six datasets, we obtained 57 divisions using different parameters. Each division was evaluated by USI-AUC criterion and modularity criterion, with evaluation results shown in Fig. 8 (sorting from small to large according to modularity).

Representative experimental results are listed in Table 8. For each community detection algorithm, we selected an optimal parameter according to the USI-AUC criterion and listed the scores from the two criteria and their corresponding rankings.

5.2.1. Comparative analysis of general trends

For the disjoint-community detection algorithms, the USI-AUC and modularity criteria exhibited positively correlated trends, indicating their joint capabilities. With increases in modularity, the USI-AUC also increased; however, according to data shown in Fig. 8 and Table 7, there were also fluctuations between the two indices mainly caused by isolated nodes. For example, the reason for the strong fluctuation in the fifth division of the lesmis dataset from Fig. 8(a) concerned the use of {11, 14, 15, 16, 31, 33, 34, 41, 43, and 68} nodes as isolated nodes, with the degrees of the nodes {11, 14, 15, 16, 33, 41, and 68} all equal to 1 and linked with nodes

Table 5. The brief description of 11 disjoint-community detection algorithms and corresponding parameters.

Algorithm	Brief description	Parameter(s)
StabilityOpt ³⁴	A community detection method using stability as an optimization criterion.	0.1, 0.5, 0.9, 1.3, 1.7, 2.1, 2.5, and 2.9.
Ronhovde ³⁵	An exceptionally accurate spin-glass-type Potts model for community detection.	0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, and 0.5.
SpectralClust ³⁵	A graph-partitioning algorithm.	6
SpectralClust2 ³⁶	A graph-partitioning algorithm for perceptual problems.	6
Reichardt ³⁷	Statistical mechanics of community detection.	0.5, 1.0, 1.5, 2.0, 2.5, and 3.0.
ModulMax1 ³⁸	A simple method that reveals the hierarchical structure of the network.	N/A
ModulMax2 ⁵	Classic Newmans modularity method.	N/A
LFK ¹⁵	A method based on local optimization of a fitness function.	0.5, 1.0, 1.5, 2.0, 2.5, and 3.0.
HSLSW ³⁹	A method based on a quality function and a fast local-expansion algorithm for uncovering communities.	0.5, 1.0, 1.5, 2.0, 2.5, and 3.0.
Danon ⁴⁰	A method that takes inhomogeneities in communities into account.	N/A
AFG ⁴¹	A method that allows for multiple-resolution screening of the modular structure.	0.5, 1.0, 1.5, 2.0, 2.5, and 3.0.

Table 6. The basic topological features of six example networks. $|V|$ and $|E|$ are the total numbers of nodes and links, respectively. $\langle k \rangle$ represents the average degree of nodes in a network, and $\langle d \rangle$ represents the average distance between nodes in a network. C and r are the clustering coefficient and assortative coefficient, respectively. H is the degree heterogeneity, defined as $H = \frac{\langle k^2 \rangle}{\langle k \rangle^2}$.

Network	$ V $	$ E $	$\langle k \rangle$	$\langle d \rangle$	C	r	H
Lesmis	77	258	6.70	2.63	0.723	-0.143	1.78
Zachary	34	78	4.59	2.41	0.606	-0.476	1.69
Football	115	613	10.66	2.51	0.403	0.162	1.01
CKM-3	246	423	3.44	4.24	0.356	0.102	1.33
PPI_Cell	127	237	3.73	4.45	0.455	0.035	1.65
Metabolic	453	2025	8.94	2.66	0.660	-0.226	4.49

of larger degree. The degrees of nodes $\{31, 34, \text{ and } 43\}$ were two or three and were linked with nodes of larger degree. Therefore, when node information was limited, regarding these nodes as isolated was reasonable. We were unsure whether these nodes were attached to the community containing large-degree nodes. These results were consistent with **Theorem 1**, indicating that USI-AUC criterion provided a reasonable evaluation of communities containing isolated nodes.

5.2.2. Comparative analysis using labeled datasets

We selected community labeled datasets Zachary and Football to analyze the two metrics. In the Zachary dataset, the original division consisted of two communities.

(a) Lesmis

(b) Zachary

(c) Football

(d) CKM-3

(e) PPI-Cell

(f) Metabolic

Fig. 8. Comparison of USI-AUC and modularity criteria. The “*” represents the score associated with the USI-AUC criterion, and the “+” represents the score associated with the modularity criterion. Each scatter diagram showing modularity is sorted from small to large and shows the corresponding points for the USI-AUC score.

Using the 11 algorithms, we obtained three types of divisions consisting of two communities from five algorithms. The experimental results are listed as Table 8.

Table 7. The scores and rankings of disjoint-community detection algorithms using USI-AUC and modularity criteria. We selected the optimal parameter allowing the USI-AUC criterion to reach the maximum value in each algorithm.

(a) Lesmis	USI-AUC	Ranking of USI- AUC	Modu- larity	Ranking of modu- larity	(b) Zachary	USI-AUC	Ranking of USI- AUC	Modu- larity	Ranking of modu- larity
StabilityOpt	0.824	8	0.533	7	StabilityOpt	0.748	9	0.392	8
Ronhovde	0.862	1	0.307	11	Ronhovde	0.780	6	0.359	9
SpectralClust1	0.817	10	0.428	10	SpectralClust1	0.783	1	0.407	5
SpectralClust2	0.817	10	0.432	9	SpectralClust2	0.705	11	0.252	11
Reichardt	0.823	9	0.537	6	Reichardt	0.778	7	0.399	6
ModulMax1	0.840	4	0.552	1	ModulMax1	0.781	2	0.419	1
ModulMax2	0.842	3	0.544	4	ModulMax2	0.781	2	0.419	1
LFK	0.835	7	0.545	3	LFK	0.742	10	0.343	10
HSLSW	0.837	5	0.511	8	HSLSW	0.778	7	0.399	6
Danon	0.836	6	0.544	4	Danon	0.781	2	0.419	1
AFG	0.849	2	0.552	1	AFG	0.781	2	0.419	1
(c) Football	USI-AUC	Ranking of USI- AUC	Modu- larity	Ranking of modu- larity	(d) CKM-3	USI-AUC	Ranking of USI- AUC	Modu- larity	Ranking of modu- larity
StabilityOpt	0.844	2	0.600	4	StabilityOpt	-	-	-	-
Ronhovde	0.816	10	0.562	9	Ronhovde	0.856	7	0.688	6
SpectralClust1	0.814	11	0.538	10	SpectralClust1	0.832	9	0.653	10
SpectralClust2	0.803	9	0.520	11	SpectralClust2	0.582	10	0.108	9
Reichardt	0.830	8	0.573	8	Reichardt	0.904	2	0.741	3
ModulMax1	0.841	3	0.604	1	ModulMax1	0.846	8	0.644	9
ModulMax2	0.845	1	0.603	2	ModulMax2	0.887	6	0.739	4
LFK	0.832	7	0.600	4	LFK	0.907	1	0.690	5
HSLSW	0.834	6	0.593	6	HSLSW	0.904	2	0.684	7
Danon	0.836	5	0.581	7	Danon	0.895	5	0.748	2
AFG	0.840	4	0.603	2	AFG	0.897	4	0.750	1
(e) PPI-Cell	USI-AUC	Ranking of USI-AUC	Modu- larity	Ranking of modu- larity	(f) Metabolic	USI-AUC	Ranking of USI- AUC	Modu- larity	Ranking of modu- larity
StabilityOpt	0.823	1	0.613	1	StabilityOpt	0.734	4	0.434	3
Ronhovde	0.822	2	0.502	10	Ronhovde	0.730	6	0.194	9
SpectralClust1	0.701	11	0.321	11	SpectralClust1	0.687	10	0.277	8
SpectralClust2	0.810	7	0.535	8	SpectralClust2	-	-	-	-
Reichardt	0.811	6	0.561	6	Reichardt	0.742	1	0.408	6
ModulMax1	0.816	3	0.604	3	ModulMax1	0.734	4	0.421	4
ModulMax2	0.813	4	0.608	2	ModulMax2	0.735	3	0.437	2
LFK	0.788	9	0.560	7	LFK	0.711	8	0.337	7
HSLSW	0.762	10	0.513	9	HSLSW	0.694	9	0.152	10
Danon	0.812	5	0.603	4	Danon	0.723	7	0.416	5
AFG	0.810	7	0.603	4	AFG	0.739	2	0.440	1

From rankings of USI-AUC and modularity results, the two metrics are almost the same based on evaluating the division of the Zachary dataset, with the original division of each scoring highest in the two indices and showing that both methods

Table 8. Comparing the evaluation of USI-AUC and modularity criteria using the labeled dataset Zachary.

Zachary	USI-AUC	Ranking of USI-AUC	Modularity	Ranking of modularity
Original division	0.730	1	0.371	1
SpectralClust1	0.657	4	0.298	4
SpectralClust2/Reichardt	0.697	3	0.360	2
LFK/HLSW	0.710	2	0.352	3

were reasonable.

The same experiment using the Football dataset (see Table 9) showed that both USI-AUC and modularity criteria were better for the divisions of the algorithms as compared with the original communities, indicating that the two indexes either exhibited evaluation errors or some communities were unlabeled. For example, although nodes representing teams in the dataset were attributed to their own leagues, there may still be teams capable of establishing relationships with teams from other leagues, resulting in potentially unmarked community structures.

Table 9. Comparison of evaluations using USI-AUC and modularity criteria on the labeled dataset Football.

Football	USI-AUC	Ranking of USI-AUC	Modularity	Ranking of modularity
Original division	0.799	4	0.554	4
StabilityOpt	0.839	1	0.597	1
Reichardt	0.822	3	0.567	3
HLSW	0.838	2	0.595	2

5.3. Evaluation of the overlapping-community detection algorithm using USI-AUC criterion

If communities overlap, USI-AUC criterion can also be used for their evaluation. For example, Fig. 1 shows that using nodes 3 and 9 as common nodes in the Zachary dataset results in a value according to USI-AUC criterion of 0.742, which is greater than the USI-AUC value associated with the original division. However, the value of EQ was 0.351, which was lower than the EQ value associated with the original division according to **Theorem 2**. As illustrated in Fig. 1, nodes 3 and 9 are active in both communities; therefore, use of the USI-AUC criterion was more reasonable.

For evaluation of the overlapping-community detection algorithm, we used the Dolphin and Football datasets and partitioned the networks using the fast overlapping community detection algorithm (FOCS)⁴⁶ and the other 11 disjoint algorithms. Evaluation results according to USI-AUC and EQ are shown in Table 10.

Use of USI-AUC criterion was able to evaluate both the disjoint and overlapping

Table 10. Rankings of disjoint- and overlapping-community detection algorithms according to the use of USI-AUC and extended modularity (EQ) criteria on the Dolphin and football datasets based on selecting the parameter optimizing USI-AUC-criterion results for both algorithms.

(a) Dolphin	USI-AUC	Ranking of USI -AUC	Modu- larity	Ranking of modu- larity	(b) Football	USI-AUC	Ranking of USI -AUC	Modu- larity	Ranking of modu- larity
FOCS(2 0.5)	0.816	1	0.361	13	FOCS(2 0.5)	0.796	13	0.484	12
FOCS(1 0.5)	0.797	2	0.376	11	FOCS(1 0.5)	0.799	12	0.418	13
StabilityOpt	0.795	4	0.523	1	StabilityOpt	0.844	2	0.600	4
Ronhovde	0.782	9	0.501	7	Ronhovde	0.816	9	0.562	9
SpectralClust1	0.785	7	0.484	8	SpectralClust1	0.814	10	0.538	10
SpectralClust2	0.733	13	0.384	10	SpectralClust2	0.803	11	0.520	11
Reichardt	0.790	6	0.518	3	Reichardt	0.830	8	0.573	8
ModulMax1	0.780	10	0.502	6	ModulMax1	0.841	3	0.604	1
ModulMax2	0.748	12	0.373	12	ModulMax2	0.845	1	0.603	2
LFK	0.791	5	0.521	2	LFK	0.832	7	0.600	4
HSLSW	0.768	11	0.473	9	HSLSW	0.834	6	0.593	6
Danon	0.785	7	0.506	5	Danon	0.836	5	0.581	7
AFG	0.797	2	0.513	4	AFG	0.840	4	0.603	2

communities. Based on the evaluation results, values associated with the USI-AUC criterion were highest when using the FOCS algorithm with parameters (2 and 0.5), indicating that this type of division was more reasonable. However, the value associated with use of USI-AUC criterion for the FOCS algorithm was lower than that observed using the other algorithms on the Football dataset, indicating that this type of overlapping division was of little significance. As for the *EQ* criterion, because of the overlapping communities, the *EQ* value was low, which was consistent with the analysis using **Theorem 2**. Therefore, *EQ* did not provide a reasonable evaluation of the overlapping-community detection algorithm.

5.4. Statistical significance test for the comparison of the USI-AUC measure with other measures

In the experiment, we use some artificial network data sets generated by LFR benchmark⁴⁷ with community labels to do comparative analyses with other evaluation measures such as precision, recall, F1-score, NMI, omega and modularity. The parameters description of LFR benchmark is listed in table 11; parameters setting of LFR benchmark to generate 4 networks with labeled overlapping communities is shown in table 12; the basic topological features of 4 generated networks are listed in table 13.

FLR1, FLR2, LFR3, LFR4 are networks generated by LFR benchmark within overlapping communities. FLR1-1, FLR2-1, FLR3-1, FLR4-1 are the disjoint community networks by adjusting common nodes to a community randomly from FLR1, FLR2, LFR3, LFR4, respectively. FLR1-2, FLR2-2, FLR3-2, FLR4-2 are adjusted networks within overlapping communities simulated as algorithm found. We can see from table 14, USI-AUC criterion has a better consistency with criteria based on

Table 11. Parameter description of LFR benchmark.

Parameter	Description
$ V $	number of nodes
$\langle k \rangle$	average degree
maxk	maximum degree
mu	mixing parameter
minc	minimum for the community sizes
maxc	maximum for the community sizes
on	number of overlapping nodes
om	number of memberships of the overlapping nodes
C	Average clustering coefficient

Table 12. Parameter setting of LFR benchmark network generation.

Network	$ V $	$\langle k \rangle$	maxk	mu	minc	maxc	on	om	C
LFR1	500	8	50	0.1	15	40	20	3	0.6
LFR2	500	8	50	0.3	15	40	20	4	0.6
LFR3	500	10	50	0.1	15	40	20	3	0.6
LFR4	500	10	50	0.3	15	40	20	4	0.6

Table 13. The basic topological features of 4 generated networks.

Network	$ V $	$ E $	$\langle k \rangle$	$\langle d \rangle$	C	r	H
LFR1	500	2028	8.112	3.897	0.629	-0.148	1.877
LFR2	500	2152	8.608	3.735	0.566	-0.063	2.060
LFR3	500	2340	9.360	3.543	0.606	-0.098	1.762
LFR4	500	2345	9.740	3.818	0.583	-0.021	1.726

labeled network. However, there exists phenomenon the modularity of disjoint community is higher than the extended modularity of standard overlapping community, which is assistant with **theorem 2**.

6. Conclusions

In this paper, we have proved that ME actually has theoretical defects in the presence of evaluation disjoint and overlapping communities. On that basis, we have proposed a method using AUC accuracy for the link-prediction method based on the USI model as an evaluation criterion (USI-AUC). Avoiding defects associated with contributions from isolated nodes to modularity and the inability to determine a best overlapping community according to ME, this criterion effectively evaluated the disjoint and overlapping communities, and has a better consistency with criteria based on labeled network. However, this criterion exhibited fluctuations, indicating that USI-AUC will not be consistent based on differences in the training or test sets, indicating that this criterion requires datasets with multiple partitions.

Table 14. Statistical significance test for the comparison of the USI-AUC measure with other evaluation measures.

Network	Precision	Recall	F1-score	NMI	Omega	Modularity	USI-AUC
FLR1	1	1	1	1	1	0.771	0.928
FLR1-1	1	0.926	0.962	0.848	0.863	0.774	0.902
FLR1-2	0.896	0.861	0.878	0.784	0.822	0.666	0.869
FLR2	1	1	1	1	1	0.575	0.825
FLR2-1	1	0.893	0.943	0.824	0.846	0.577	0.805
FLR2-2	0.896	0.866	0.881	0.769	0.765	0.448	0.760
FLR3	1	1	1	1	1	0.711	0.931
FLR3-1	1	0.926	0.962	0.855	0.871	0.712	0.869
FLR3-2	0.896	0.861	0.878	0.744	0.779	0.580	0.861
FLR4	1	1	1	1	1	0.586	0.822
FLR4-1	1	0.893	0.943	0.684	0.686	0.587	0.805
FLR4-2	0.896	0.866	0.881	0.681	0.731	0.465	0.760

In this paper, we use an USI-AUC-evaluation index for the 0-order set, which can be extended to a 1-order set or a higher-order set. Our results also indicated that the USI model was suitable for weighted networks; therefore, the USI-AUC index is applicable for the evaluation of community detection in weighted networks. However, for directed networks, the USI model was not directly applied. Future work lies in generalizing the use of the USI model and USI-AUC criterion in directed networks. Moreover, we will also try to apply current HPC techniques (e.g., cloud computing⁴⁸) to meet the challenges of massive networks.

Acknowledgments

We acknowledge Xi Tan and Wei Liu for their inspirations. This work was partially supported by the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (No. 61521003), and National Natural Science Foundation of China (No. 61601513).

References

1. M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12):7821–6, 2002.
2. Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2009.
3. Long Cheng, Spyros Kotoulas, Tomas E Ward, and Georgios Theodoropoulos. Robust and skew-resistant parallel joins in shared-nothing systems. In *ACM International Conference on Conference on Information and Knowledge Management*, pages 1399–1408, 2014.
4. Zhenping Li, Rui Sheng Wang, Shihua Zhang, and Xiang Sun Zhang. Modularity and community detection in bipartite networks. *Computer Science*, 2015.
5. M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E Statistical Nonlinear & Soft Matter Physics*, 69(6):066133–066133, 2004.

6. M. E. J. Newman. Community detection in networks: Modularity optimization and maximum likelihood are equivalent. 2016.
7. R. Langone, C. Alzate, and J. A. K. Suykens. Kernel spectral clustering for community detection in complex networks. *Proc of the IEEE World Congress on Computational Intelligence*, 20:1–8, 2012.
8. Sarah Habashi, Nagia M. Ghanem, and Mohamed A. Ismail. Enhanced community detection in social networks using active spectral clustering. In *The ACM Symposium*, pages 1178–1181, 2016.
9. Qing Cheng, Zhong Liu, Jincai Huang, and Cheng Zhu. Hierarchical clustering based on hyper-edge similarity for community detection. In *Ieee/wic/acm International Conferences on Web Intelligence and Intelligent Agent Technology*, pages 238–242, 2012.
10. U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E Statistical Nonlinear & Soft Matter Physics*, 76(3 Pt 2):–, 2007.
11. X. Liu and T. Murata. Advanced modularity-specialized label propagation algorithm for detecting communities in networks. *Physica A Statistical Mechanics & Its Applications*, 389(7):1493–1500, 2010.
12. Gergely Palla, Imre Dernyi, Ills Farkas, and Tams Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–8, 2005.
13. Steve Gregory. Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 12(10):2011–2024, 2009.
14. Yong Yeol Ahn, James P. Bagrow, and Sune Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–4, 2010.
15. Andrea Lancichinetti, Santo Fortunato, and Jnos Kertsz. Detecting the overlapping and hierarchical community structure of complex networks. *New Journal of Physics*, 11(3):19–44, 2009.
16. T Nepusz, A Petrczi, L Ngyessy, and F Bazs. Fuzzy communities and the concept of bridgeness in complex networks. *Physical Review E*, 77(1 Pt 2):119–136, 2008.
17. L. M. Collins and C. W. Dent. Omega: A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions. *Multivariate Behavioral Research*, 23(2):231–242, 1988.
18. Steve Gregory. Fuzzy overlapping communities in networks. *Journal of Statistical Mechanics Theory & Experiment*, 2(2):P02017, 2011.
19. Clara Pizzuti. Ga-net: A genetic algorithm for community detection in social networks. In *Parallel Problem Solving From Nature - PPSN X, International Conference Dortmund, Germany, September 13-17, 2008, Proceedings*, pages 1081–1090, 2008.
20. V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri. Extending modularity definition for directed graphs with overlapping communities. 2008.
21. Huawei Shen, Xueqi Cheng, Kai Cai, and Mao Bin Hu. Detect overlapping and hierarchical community structure in networks. *Physica A Statistical Mechanics & Its Applications*, 388(8):1706–1712, 2009.
22. Hua Wei Shen, Xue Qi Cheng, and Jia Feng Guo. Quantifying and identifying the overlapping community structure in networks. *Journal of Statistical Mechanics Theory & Experiment*, 2009(7):07042, 2009.
23. Wayne W. Zachary. An information flow model for conflict and fission in small groups1. *Journal of Anthropological Research*, 33(4):473, 1977.
24. Jierui Xie, Stephen Kelley, and Boleslaw K. Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm Computing Surveys*, 45(4):115–123, 2013.

25. Leon Danon, Albert Dazguilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics Theory & Experiment*, 2005(09):09008, 2005.
26. Jorge Valverde-Rebaza and Alneu De Andrade Lopes. Exploiting behaviors of communities of twitter users for link prediction. *Social Network Analysis and Mining*, 3(4):1063–1074, 2013.
27. Jorge Valverde-Rebaza, Alan Valejo, Lilian Berton, Thiago De Paulo Faleiros, and Alneu De Andrade Lopes. A naive bayes model based on overlapping groups for link prediction in online social networks. In *The Acm/sigapp Symposium on Applied Computing*, pages 1136–1141, 2015.
28. Sucheta Soundarajan and John Hopcroft. Using community information to improve the precision of link prediction methods. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 607–608, 2012.
29. Clauset Aaron, Moore Cristopher, and M E J Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.
30. Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.
31. E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9(5):1981–2014, 2008.
32. J. A. Hanley and B. J. Mcneil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.
33. David Lusseau, Karsten Schneider, Oliver J. Boisseau, Patti Haase, Elisabeth Slooten, and Steve M. Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology & Sociobiology*, 54(4):396–405, 2003.
34. Erwan Le Martelot and Chris Hankin. Multi-scale community detection using stability as optimisation criterion in a greedy algorithm. 2011.
35. P Ronhovde and Z Nussinov. Local resolution-limit-free potts model for community detection. *Physical Review E Statistical Nonlinear & Soft Matter Physics*, 81(2):387–395, 2010.
36. P. Hespanha. An efficient matlab algorithm for graph partitioning. 2004.
37. J Reichardt and S Bornholdt. Statistical mechanics of community detection. *Physical Review E Statistical Nonlinear & Soft Matter Physics*, 74(2):–, 2006.
38. Vincent D. Blondel, Jean Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of community hierarchies in large networks. *J Stat Mech*, abs/0803.0476, 2008.
39. Jianbin Huang, Heli Sun, Yaguang Liu, Qinbao Song, and Tim Weninger. Towards online multiresolution community detection in large-scale networks. *Plos One*, 6(8):492–492, 2011.
40. Leon Danon, Albert Diaz-Guilera, and Alex Arenas. Effect of size heterogeneity on community identification in complex networks. *Journal of Statistical Mechanics Theory & Experiment*, 2006(11):11010, 2006.
41. Alex Arenas, Alberto Fernandez, and Sergio Gomez. Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics*, 10(5):4656–4658, 2008.
42. Donald E. Knuth. The stanford graphbase: a platform for combinatorial computing. In *Acm/sigact-Siam Symposium on Discrete Algorithms, 25-27 January 1993, Austin, Texas*, pages 41–43, 1993.
43. James Coleman and Herbert Menzel. The diffusion of innovation among physicians. *Sociometry*, 20(20):253–269, 1977.

- 44. Eric D. Kolaczyk. *Statistical Analysis of Network Data*. Springer New York, 2009.
- 45. J Duch and A Arenas. Community detection in complex networks using extremal optimization. *Physical Review E Statistical Nonlinear & Soft Matter Physics*, 72(2 Pt 2):986–1023, 2005.
- 46. Sanghamitra Bandyopadhyay, Garisha Chowdhary, and Debarka Sengupta. Focs: Fast overlapped community search. *IEEE Transactions on Knowledge & Data Engineering*, 27(11):2974–2985, 2015.
- 47. A Lancichinetti, S Fortunato, and F Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E Statistical Nonlinear & Soft Matter Physics*, 78(2):046110, 2008.
- 48. Long Cheng and Spyros Kotoulas. Efficient skew handling for outer joins in a cloud computing environment. *IEEE Transactions on Cloud Computing*, pages 1–1, 2015.