

Community Detection Using Random-walk Similarity and Application to Image Clustering

Makoto Okuda* Shin'ichi Satoh*† Shoichiro Iwasawa*
Shunsuke Yoshida* Yutaka Kidawara* Yoichi Sato‡

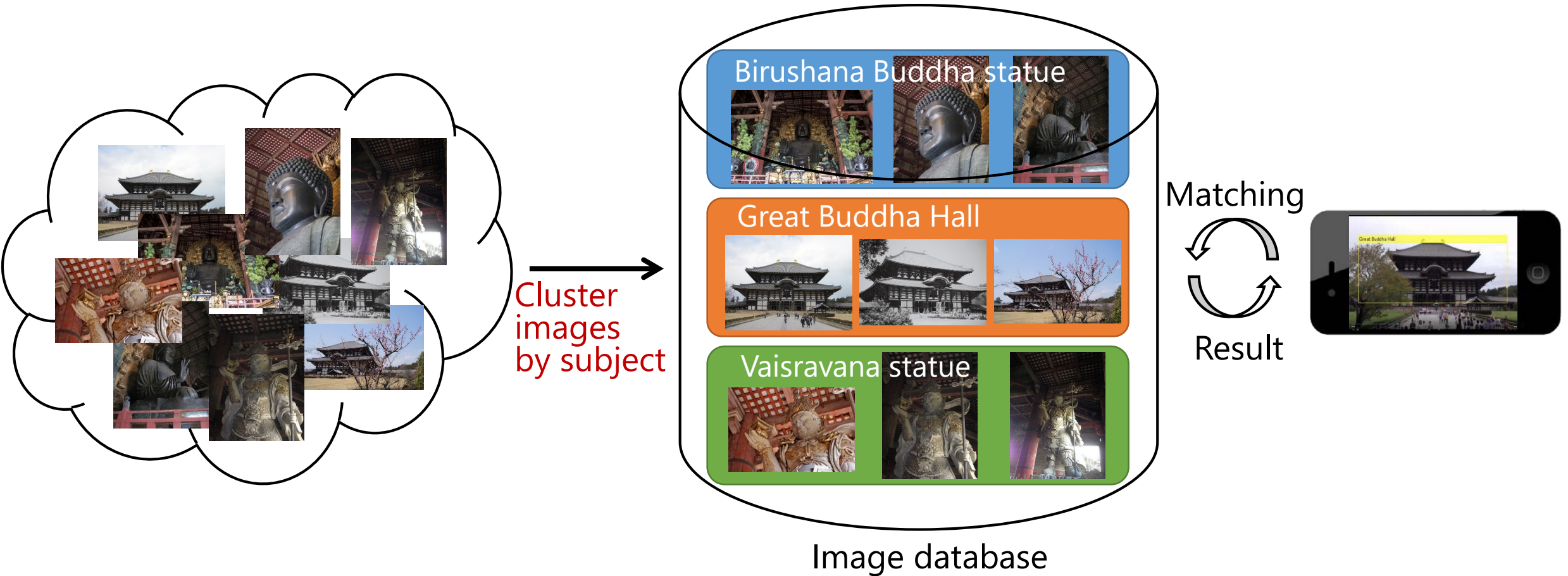
* National Institute of Information and Communications Technology (NICT)

† National Institute of Informatics

‡ The University of Tokyo

September 19, 2017

Building image database for recognition of tourist attractions



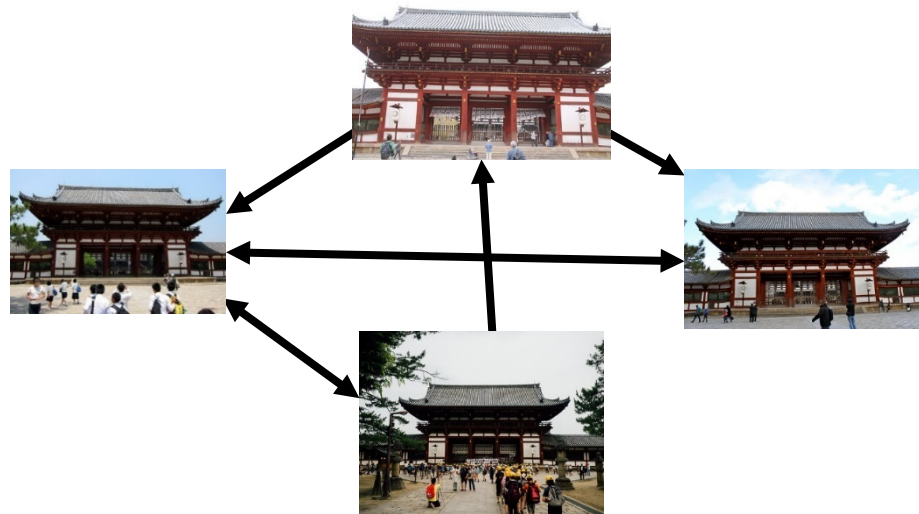
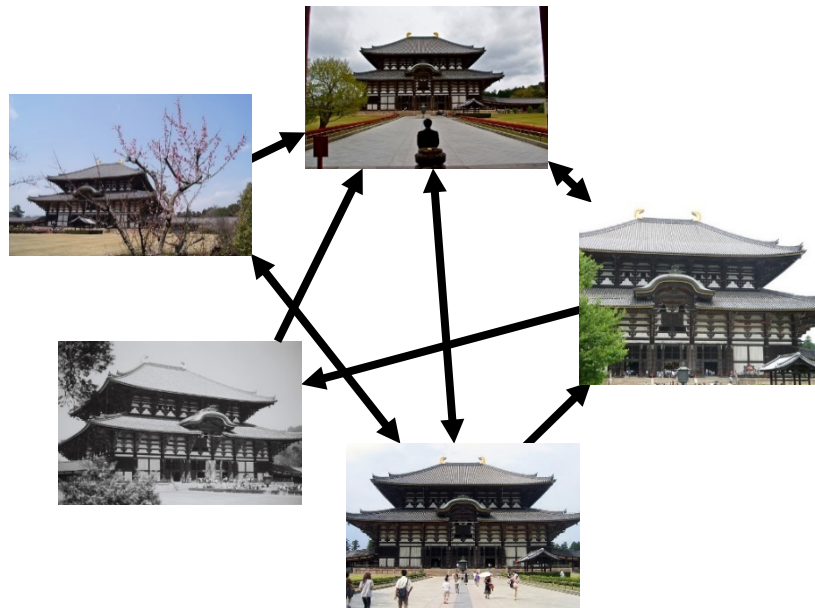
Clustering images by subject [S. Agarwall *et al.*, 2009]

1. Generate a match graph

- Nodes: Images
- Edges: SIFT keypoints matching between the images



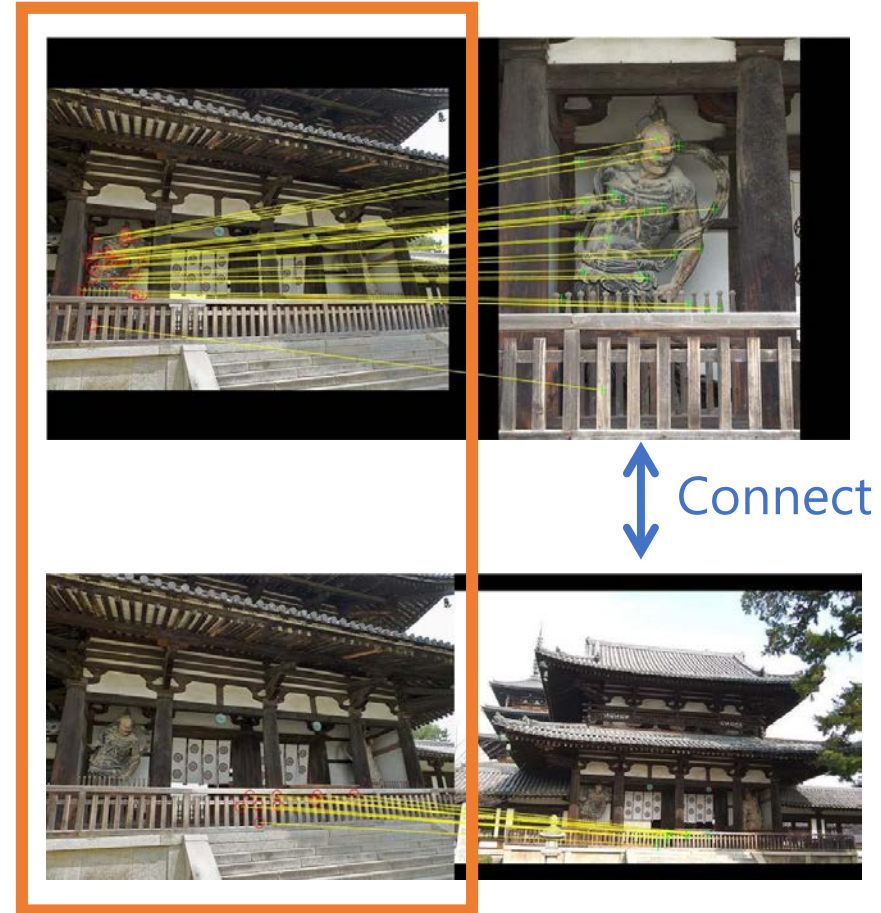
2. Detect connected components



Undesirable connection between different subjects

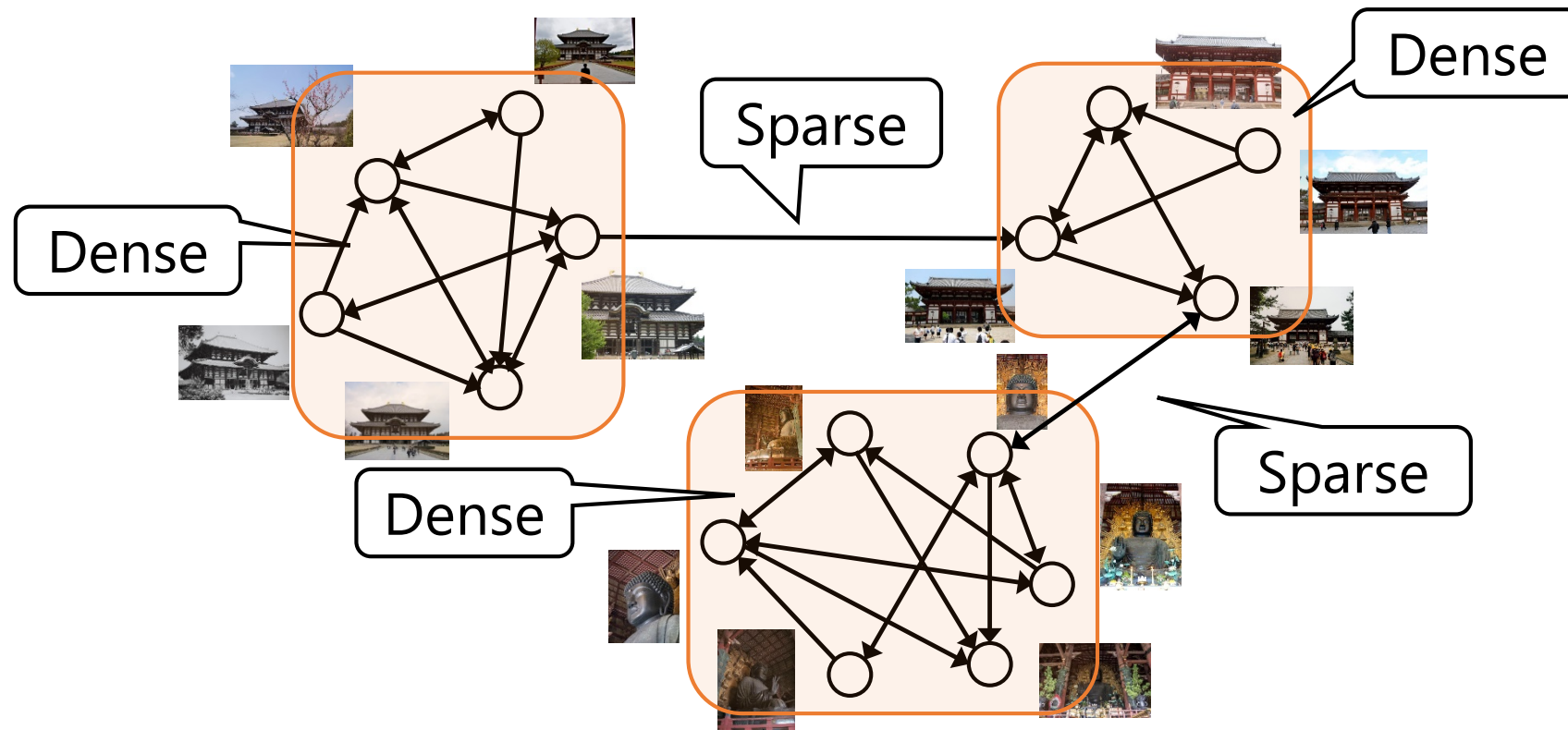


Mismatches between SIFT keypoints



Bridges by image including two subjects

Undesirable connected component



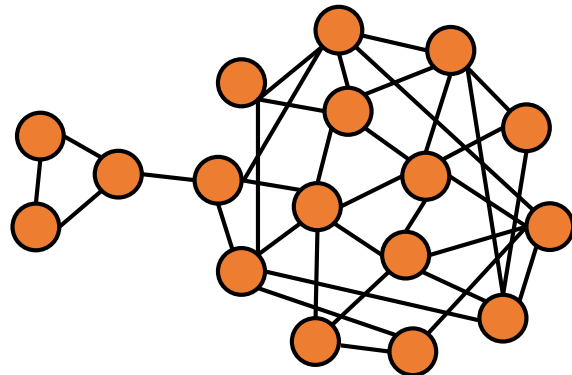
Obtain subgraphs with dense edge connections → **Community detection**

Limitation of previous community detection methods

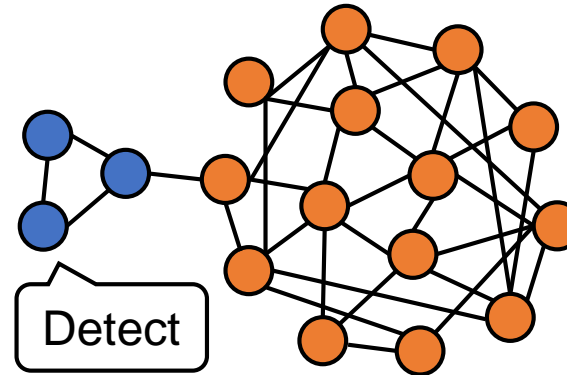
- Can not detect relatively small communities
 - Detect communities that maximize modularity Q
 - Q : A quality index for the partition of a graph into communities
 - The problem caused by the characteristic of Q



Seek a method which does not rely on Q



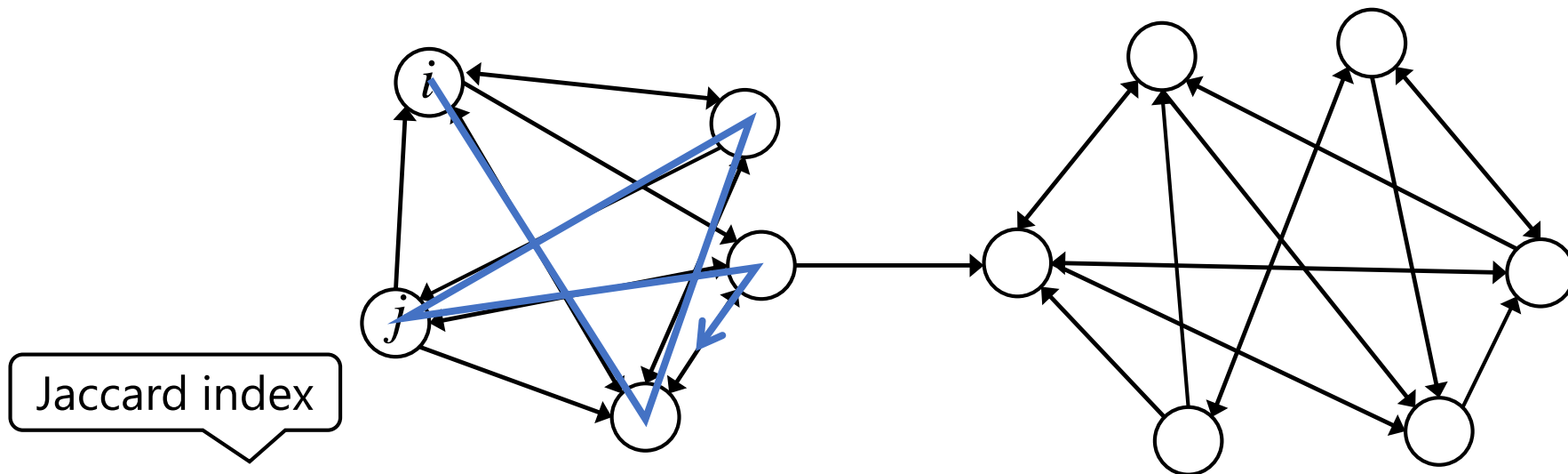
Previous methods



A proposed method

The basic idea

A random walker starting from each node in the same community passes similar nodes in a few steps.



Jaccard index

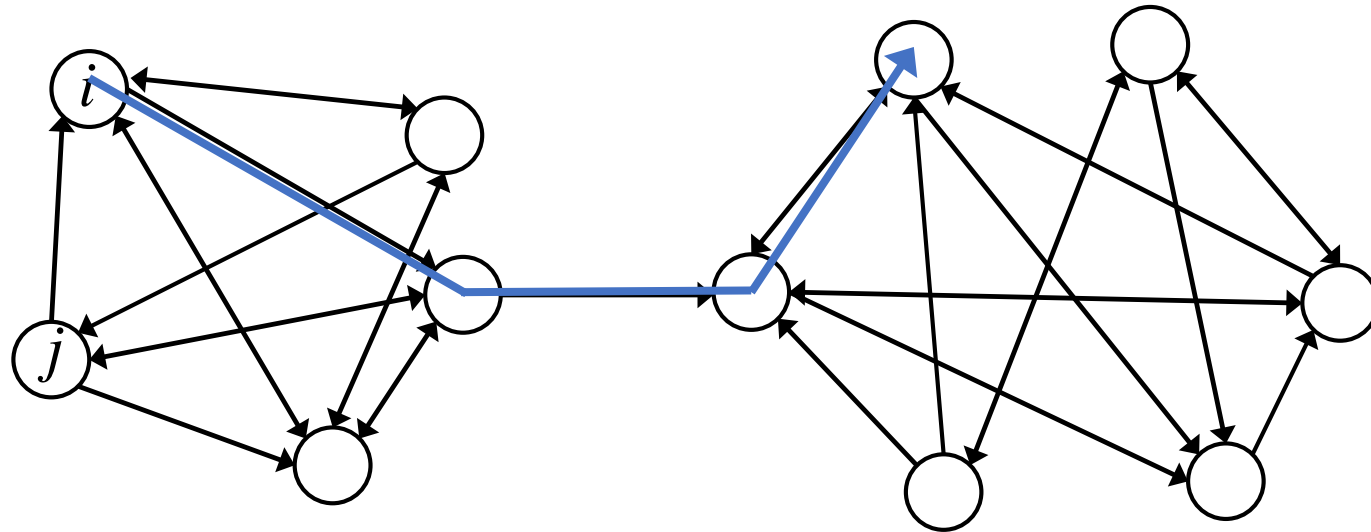
$$sim = \frac{|S_i \cap S_j|}{|S_i \cup S_j|} \geq threshold \rightarrow \text{Cluster node } i \text{ and } j \text{ into the same community}$$

S_i : Set of nodes passed by a random walker who started from the node i

S_j : Set of nodes passed by a random walker who started from the node j

Restrain effects of accidental random walks

A random walker sometimes soon leaves for the other communities



1. Execute random walks n times starting from the same node
2. Ignore the nodes that were passed very few times

Why can relatively small communities be found?

Small community

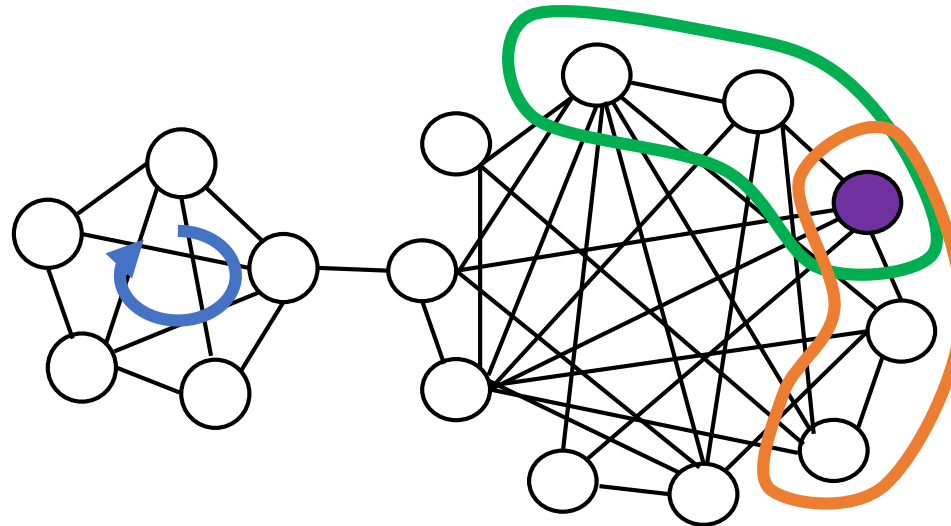
Random walkers do not easily leave for other communities **in a few steps**



Random walk similarities between the nodes in a small community become large



The small community is detected



Large community

Random walkers visit only a few of nodes **in a few steps**



Many small communities are built



Small communities which have common nodes are repeatedly united



A large community is detected

Experiments

1. Obtained the tourist attraction images from Flickr

- *Todaiji*: 4,015
- *Toshogu*: 3,808
- *Horyuji*: 1,102

2. Made three match graphs

3. Obtained 18 connected components

4. Applied the community detection methods to each connected component

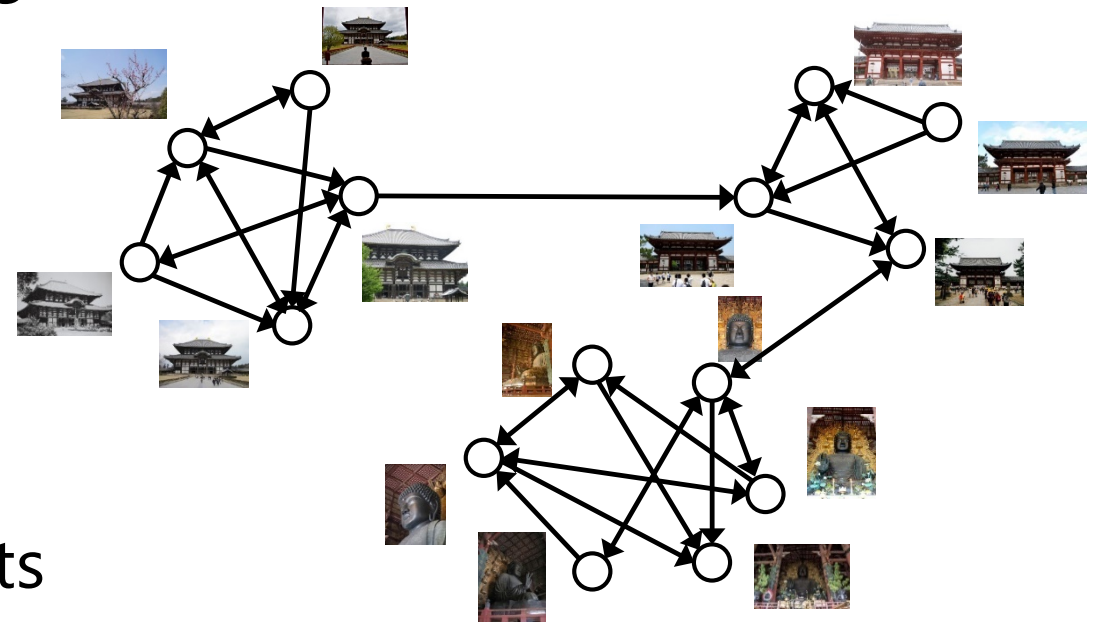








Image of a connected component

Communities detected from a connected component

Major subject	Typical image	The number of images	J. Reichardt <i>et al.</i> , 2006	M. Rosvall <i>et al.</i> , 2009	X. Fu <i>et al.</i> , 2013	P.D. Meo <i>et al.</i> , 2014	Ours
Great Buddha Hall		461	✓	✓	✓	✓	✓
Birushana Buddha statue		248	✓	✓	✓	✓	✓
Chumon Gate		100	✓	✓	✓	✓	✓
Kokuzo Bosatsu statue		54	✓	✓		✓	✓
Vaisravana statue		27	Could not detect small communities				✓
Ungyo statue		19	✓				✓

Statistical results of community detection

	J. Reichardt <i>et al.</i>, 2006	M. Rosvall <i>et al.</i>, 2009	X. Fu <i>et al.</i>, 2013	P.D. Meo <i>et al.</i>, 2014	Ours
The number of detected subjects	19	15	17	12	22
Mean of 18 global purities	0.887	0.940	0.814	0.890	0.905
Mean of 18 inverse purities	0.854	0.612	0.749	0.227	0.905
Mean of 18 F-measures	0.860	0.704	0.758	0.339	0.902

Global purity: Mean of rates of major category elements in the clusters

Inverse purity: Mean of recalls of major category elements in the clusters

F-measure: Harmonic mean of global purity and inverse purity

Conclusions

- Developed a new community detection method using random walk
- The proposed method can detect small and large communities simultaneously
- Experiments using tourist attraction images demonstrated the advantage of the proposed method over the previous methods
- Future works
 - Evaluation using larger datasets and those of other types