上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

Research on Community Detection and Community Correlation of Academic Large Data

Zhang Xilun
Shanghai Jiao Tong University
zhangxilun@sjtu.edu.cn

What is a community?

The community is usually defined as a set of internal has close contact with a group of nodes and the nature of the common features.There may be overlapping between different communities.

A difficult problem now is the discovery of communities in the network, known as community testing, as a fundamental issue of network science, which has attracted a great deal of attention in the past few decades.

BIGCLAM

There are many models attempting to solve the problem. One of the most famous is BIGCLAM.

In other models, the authors are based on the assumption that the community between the internal connection density is far greater than the density of the area between the overlapping community.But in this model, the author carried out the opposite assumption.

MMSB

Another famous model is called MMSB model, whose full name is A, Mixed-Membership, Stochastic, Blockmodel.

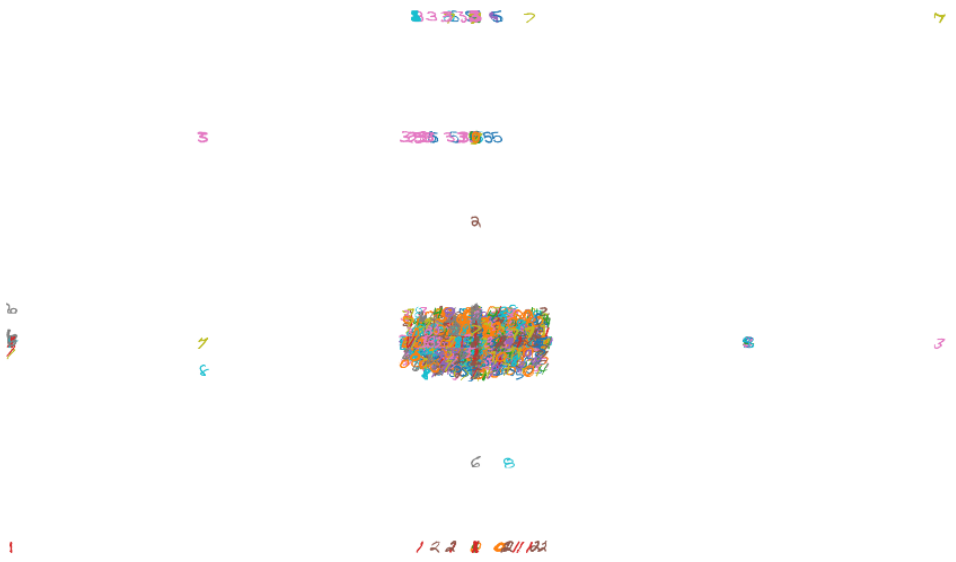
The main idea is to construct a membership vector $\theta_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{iK})$ of node i in the community. For node pairs (i, j) in the network, select community indicators $z_{i \to j}$ and $z_{i \leftarrow j}$ pointing to one of the K communities. When both are k, the probability of connection for the two nodes is β_k.

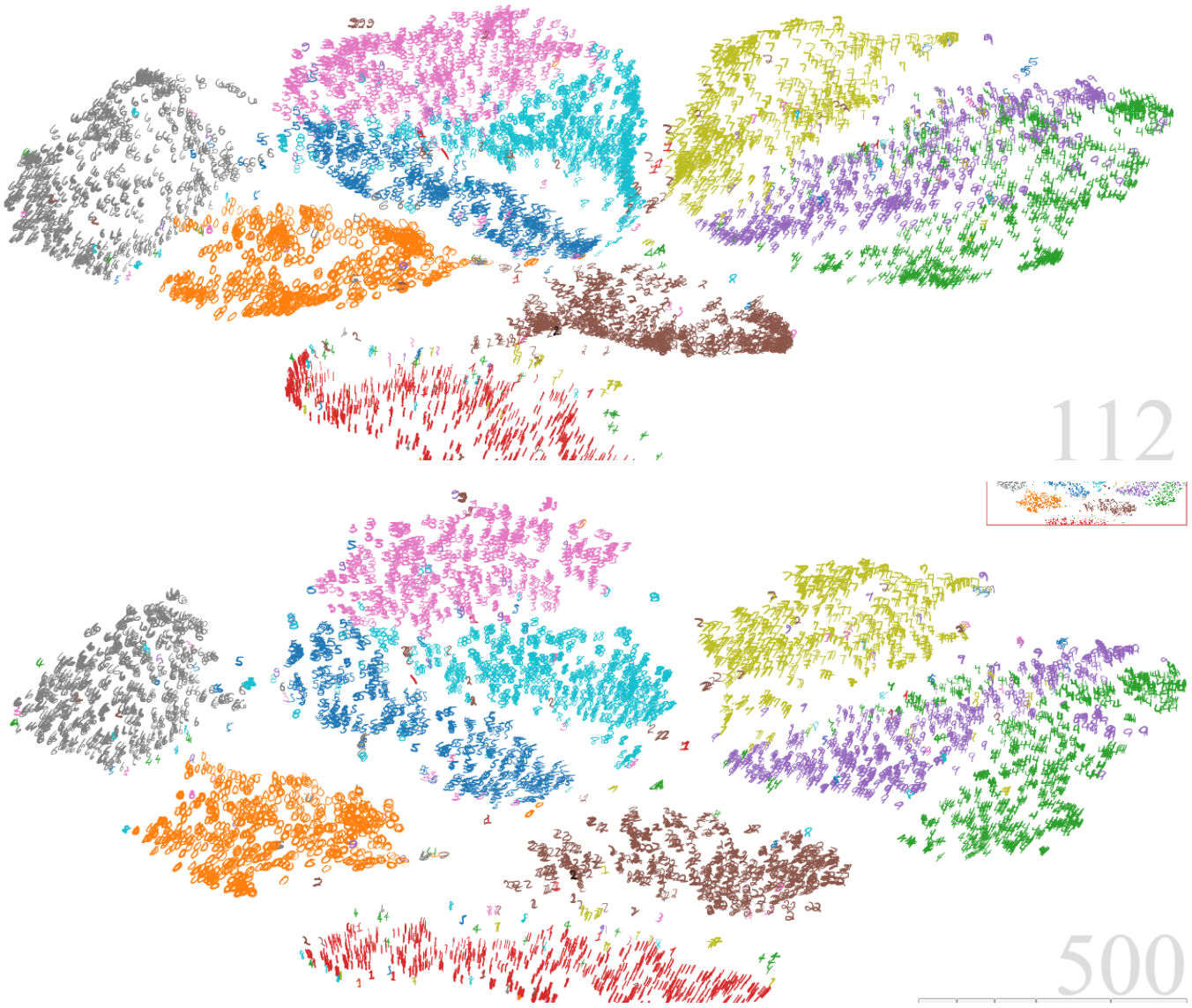
T-SNE theory

T-distributed random adjacent embedding (t-SNE) is a machine learning algorithm for dimensionality reduction. It is a nonlinear dimensionality reduction technique, which is especially suitable for embedding high-dimensional data into 2D or 3D space. Visualize in scatter charts. Specifically, it models each high-dimensional object by two-dimensional or three-dimensional points, so that similar objects are modeled by nearby points, and non-similar objects are modeled by remote points.

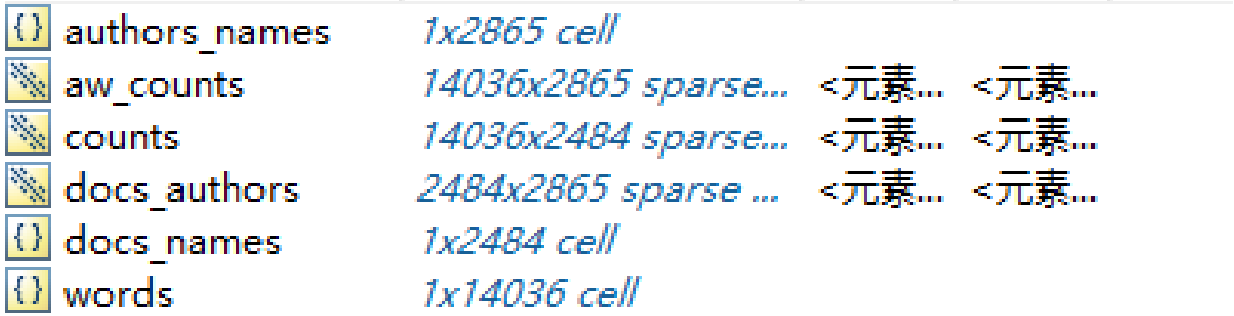
Simulation

The main principle of the preparation of the front of the map is that we first sort each of the processed images in accordance with the time, and then use JavaScript code to control each picture in a certain time interval appears on the page, the specific effect shown in the figure.




112
500

Academic article data



words: Words that appear in academic articles.

docs_name: the name of the article, the name of the format for the "year / article name."

authors_name: the author's name.

docs_authors: it is the number of articles x the number of sparse binary matrices. If an article belongs to an author, the element value at the corresponding location is 1, otherwise it is 0.

counts: it is the number of words x the number of articles count matrix.

aw_counts: it is a number of words x the number of authors of the counting matrix.

Conclusion and future prospects

In the process we do not see the changes of intermediate variables, such as the BIGCLAM model between community and node weights and MMSB membership vector, can only see the scoring algorithm finally prediction system, which is not conducive to our code debugging and modification of the algorithm. So we want to make the program when the output value of the intermediate variables, but due to the large complex network, so that all the values of all output variables is not realistic, so we hope to be able to let the node can program every fixed output intermediate results, this is our expectation for the future of the program.

Thank you!