INFORMATION THEORY AND INFERENCE

Final Project

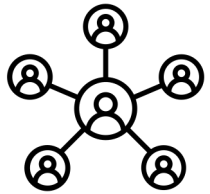
Missing and spurious interactions and the reconstruction of complex networks

Prof. Michele Allegra

Asal Rangrasiasl - 2046832

20/11/2023

CONTENTS



- INTRODUCTION
- AIM OF THE PROJECT
- METHODS
- RESULTS
- CONCLUSION & QUESTIONS

Reference Paper

Missing and spurious interactions and the reconstruction of complex networks[1]

[1]Guimerà, R., & Sales-Pardo, M. (2009). Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, *106*(52), 22073-22078.

Missing and spurious interactions and the reconstruction of complex networks

Roger Guimerà[a,1] **and Marta Sales-Pardo**[a,b,c]

[a]Department of Chemical and Biological Engineering and [b]Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL 60208; and [c]Northwestern University Clinical and Translational Science Institute, Northwestern University, Chicago, IL 60611

Edited by H. Eugene Stanley, Boston University, Boston, MA, and approved October 26, 2009 (received for review July 24, 2009)

Network analysis is currently used in a myriad of contexts, from identifying potential drug targets to predicting the spread of epidemics and designing vaccination strategies and from finding friends to uncovering criminal activity. Despite the promise of the network approach, the reliability of network data is a source of great concern in all fields where complex networks are studied. Here, we present a general mathematical and computational framework to deal with the problem of data reliability in complex networks. In particular, we are able to reliably identify both missing and spurious interactions in noisy network observations. Remarkably, our approach also enables us to obtain, from those noisy observations, network reconstructions that yield estimates of the true network properties that are more accurate than those provided by the observations themselves. Our approach has the potential to guide experiments, to better characterize network data sets, and to drive new discoveries.

data reliability | block model | modularity | node roles | Bayesian inference

The structure of the network of interactions between the units of a system affects the system's dynamics and conveys information about the functional needs of the system, its evolution, and the role of individual units. For these reasons, network analysis has become a cornerstone of fields as diverse as systems biology and sociology (1). Unfortunately, the reliability of network data is often a source of concern. In systems biology, high-throughput technologies hold the promise to uncover the intricate processes within the cell but are also reportedly inaccurate. Protein interaction data provide, arguably, the most blatant example of data inaccuracy: In 2002, a systematic comparison of several high-throughput methods to a reference high-quality data set showed that these methods have accuracies below 20% (2). Additionally, different methods result in networks that have different topological properties (3), and the coverage of real interactomes is very limited: 80% of the interactome of yeast (3) and 99.7% of the human interactome (4, 5) are still unknown.

In the social sciences, missing data due to individual nonresponse and dropout (6), informant inaccuracy (7), and sampling biases (8) are also pervasive. Simulation studies have established that these inaccuracies can lead to fundamentally wrong estimates of network properties and to misleading conclusions (8), which is particularly worrisome at a time when social network analysis is being used for finding new friends and partners, singling out key individuals in organizations, and identifying criminal activity.

Despite these concerns, the issue of network reliability has only been addressed in a field-by-field basis [for example, to deal with protein–protein interactions (9, 10) or to take into account informant inaccuracy in social networks (7)], and in studies that only address parts of the problem [for example, to detect missing interactions (11)]. Therefore, a general framework to deal with the problem of data reliability in complex networks is lacking. Here, we develop such a framework. Specifically, we show that within our framework we can reliably (*i*) identify false negatives (missing interactions) and false positives (spurious interactions) as well as (*ii*) generate, from a single observed network, a reconstructed network whose properties (clustering coefficient, modularity, assortativity, epidemic spreading threshold, and synchronizability, among others) are closer to the "true" underlying network than those of the observed network itself. We show that our approach outperforms previous attempts to predict missing and spurious interactions, and illustrate the potential of our method by applying it to a protein interaction network of yeast (12). We end by discussing how our approach will help to guide experiments and new discoveries, and to better characterize important data sets.

General Reliability Formalism

Consider an observed network with adjacency matrix A^O; $A^O_{ij} = 1$ if nodes i and j are connected and 0 otherwise. We assume that this observed network is a realization of an underlying probabilistic model, either because the network itself is the result of a stochastic process, because the measurement has uncertainty, or both (7).* Let us call $\mathcal{M}$ the set of generative models that could conceivably give rise to the observed network, and $p(M|A^O)$ the probability that $M \in \mathcal{M}$ is the model that gave rise to the observation A^O. If we could get a new observation of the network, the outcome would in general be different from A^O; our best estimate for the probability $p(X = x)$ for an arbitrary network property X is

$$p(X = x|A^O) = \int_{\mathcal{M}} dM\, p(X = x|M)p(M|A^O), \qquad [1]$$

where $p(X = x|M)$ is the probability that $X = x$ in a network generated with model M. Using Bayes theorem, we can rewrite Eq. **1** as

$$p(X = x|A^O) = \frac{\int_{\mathcal{M}} dM\, p(X = x|M)p(A^O|M)p(M)}{\int_{\mathcal{M}} dM'\, p(A^O|M')p(M')}, \qquad [2]$$

where $p(A^O|M)$ is the probability that model M gives rise to A^O among all possible adjacency matrices, and $p(M)$ is the a priori probability that model M is the correct one. We call $p(X = x|A^O)$ the reliability of the $X = x$ measurement.

Stochastic Block Models

Given the generality of these arguments, the key to good estimates of reliability is to identify sets of models that are general, empirically grounded, and analytically or computationally tractable. Here, we focus on the family $\mathcal{M}_{\mathrm{BM}}$ of stochastic block models (13, 14). In a stochastic block model, nodes are partitioned into

Author contributions: R.G. and M.S.-P. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

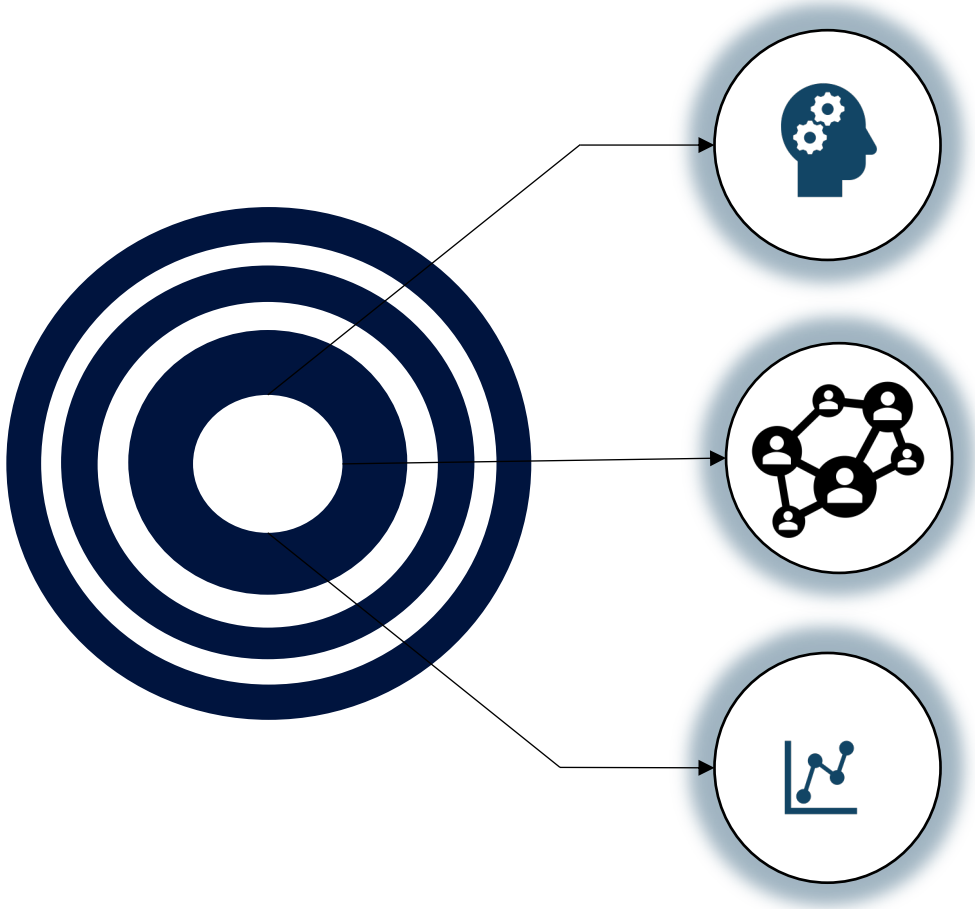
This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: rguimera@northwestern.edu.

*For simplicity, in this article we use language that is consistent with a situation in which a true network exists but is obscured by the inaccuracies of the observation process. Thus, we talk about the "true" network, which has no "errors," and about "observed" networks, which have "errors." However, the formalism is valid even if the network is itself the outcome of a stochastic process.

This article contains supporting information online at www.pnas.org/cgi/content/full/0908366106/DCSupplemental.

AIM OF THE PROJECT



Building a framework to analyse reliability of network

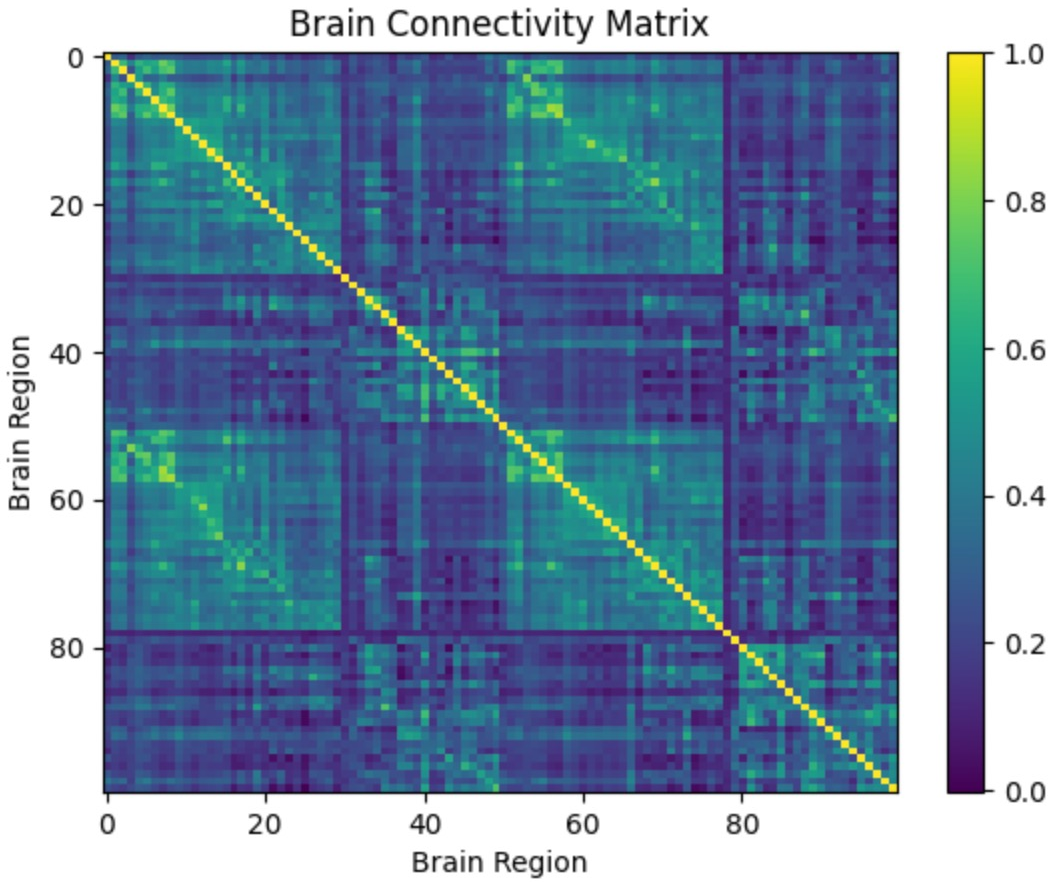
Identify spurious and missing interactions

The framework can be helpful in reconstruction of networks

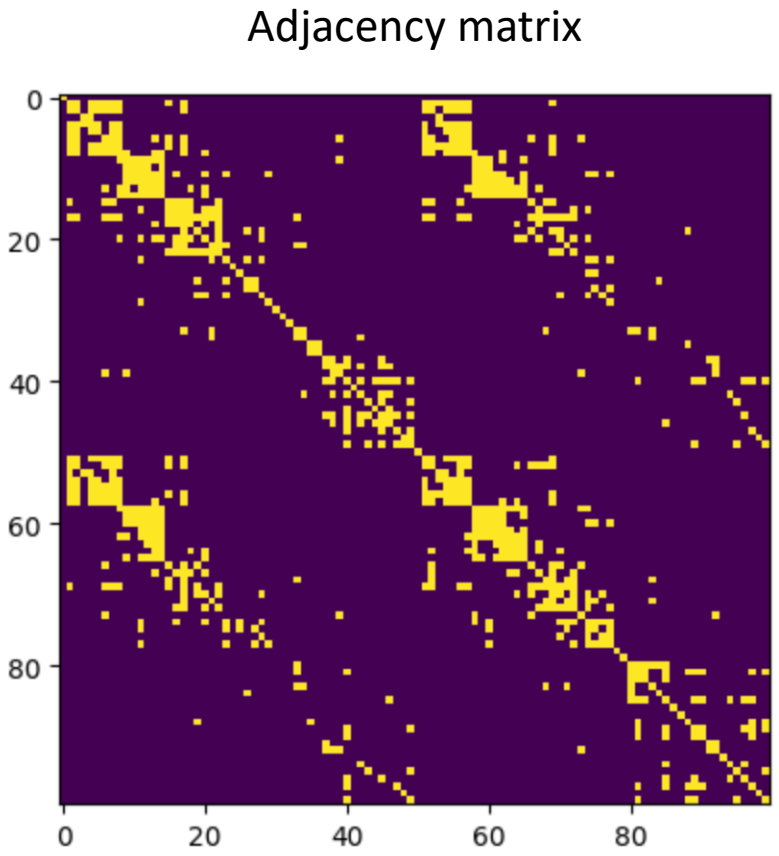
MATERIAL AND METHODS

Understanding The Context	Retrieving The Data	Dataset Preparation	Building Models and Training	CONCLUSION
The reliability of network data is a source of great concern in all fields	FC-Brain dataset	Consider an observed network with adjacency matrix	stochastic Block Models & Calculate reliability with using metropolis algorithm	Network Reconstruction

FC-Brain dataset

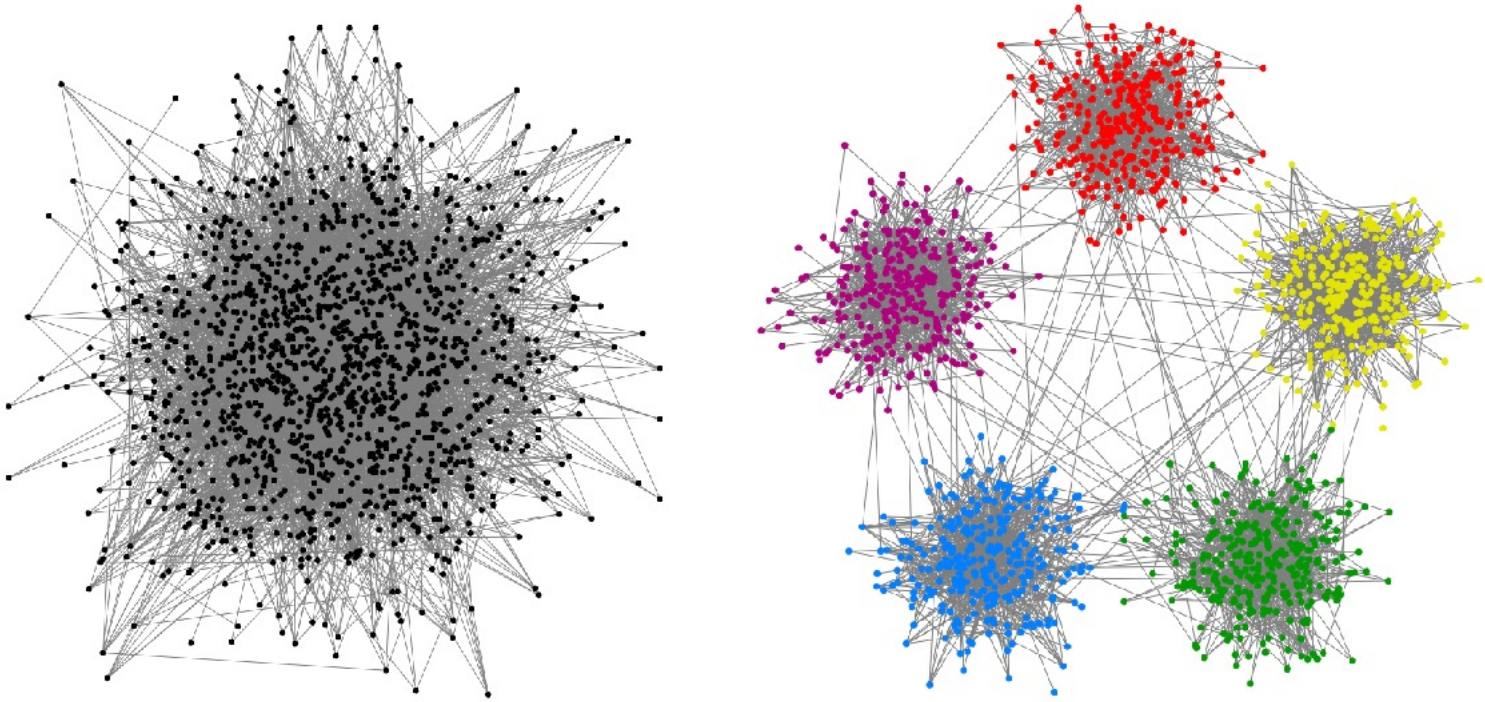


Threshold = 0.5



Stochastic Block Model

The stochastic block model (SBM) is a probabilistic model used in network analysis and community detection. It is a way to describe the generation of random graphs where nodes are partitioned into groups or "blocks," and the probability of edges between nodes depends on the block to which each node belongs. In other words, it's a generative model for random graphs with a modular structure.



Link Reliability

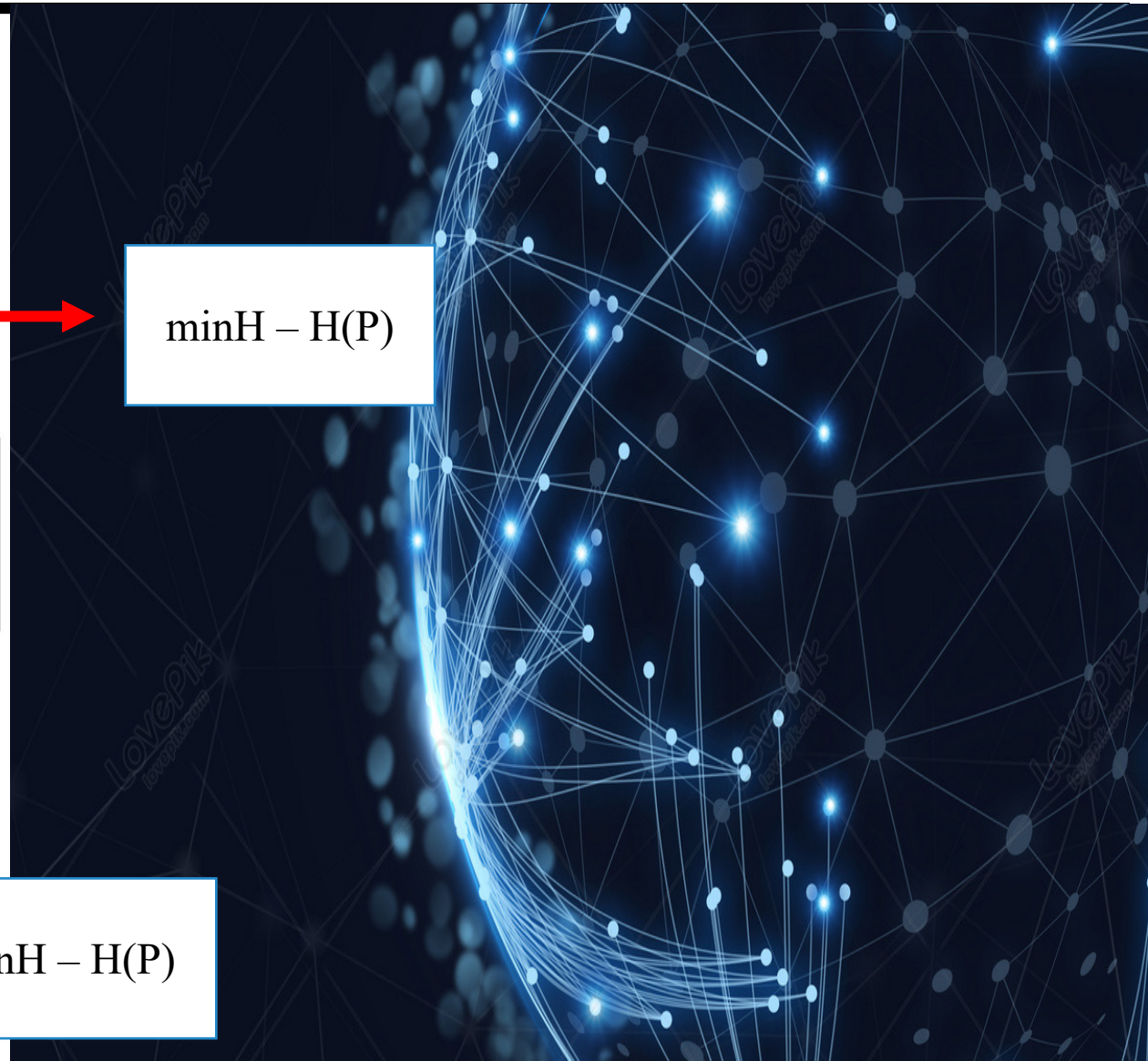
$$R_{ij}^{L} = \frac{1}{Z} \sum_{P \in \mathcal{P}} \left(\frac{l_{\sigma_i \sigma_j}^{O} + 1}{r_{\sigma_i \sigma_j} + 2} \right) \exp[-\mathcal{H}(P)]$$

minH – H(P)

$$\mathcal{H}(P) = \sum_{\alpha \leq \beta} \left[\ln(r_{\alpha\beta} + 1) + \ln \binom{r_{\alpha\beta}}{l_{\alpha\beta}^{O}} \right]$$

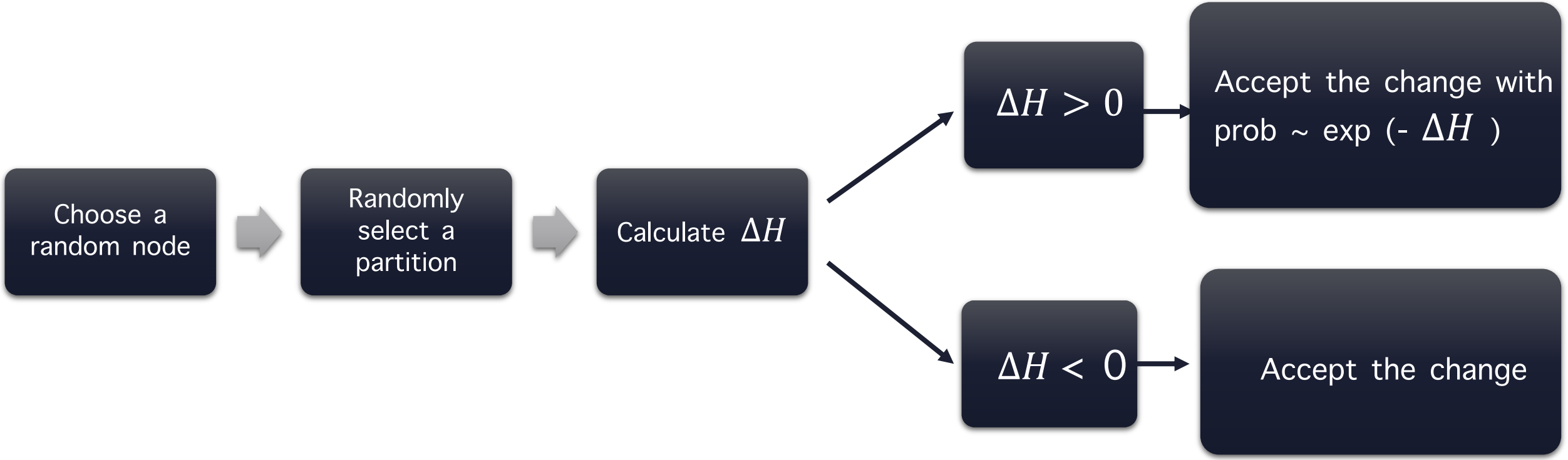
$$Z = \sum_{P \in \mathcal{P}} \exp[-\mathcal{H}(P)]$$

minH – H(P)



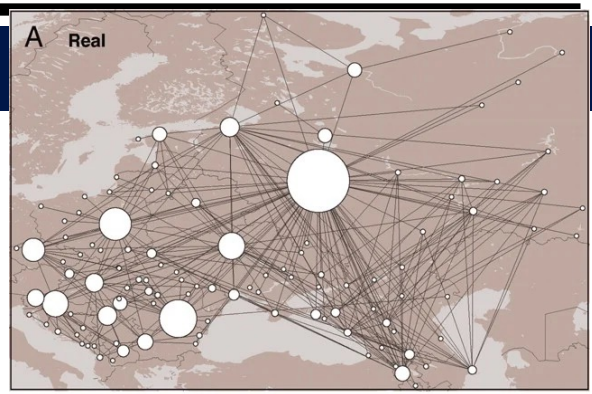
Metropolis Algorithm

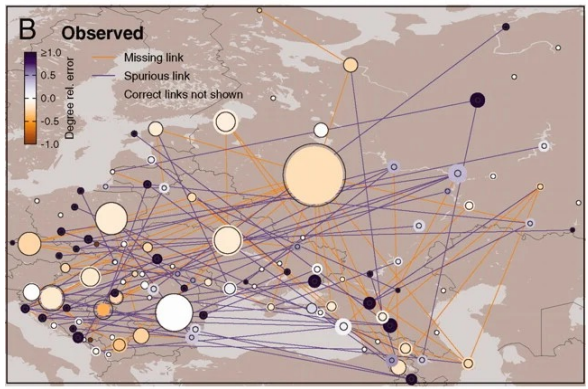
The Metropolis algorithm is a Markov chain Monte Carlo (MCMC) method used for generating random samples from a probability distribution.
In practice, it is not possible to sum over all partitions even for small networks.
we can use the Metropolis algorithm to correctly sample relevant partitions (that is, partitions that significantly contribute to the sum) and obtain estimates for the link reliability.

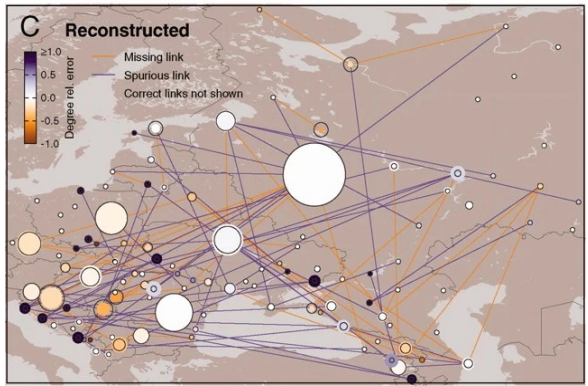


Network reconstruction

- The challenge of reconstructing a true network A-T based on an observed network A-O
- The main difficulty is that we typically don't know in advance how many missing and spurious interactions exist in the network
- To address this challenge, the authors propose a measure called "network reliability”

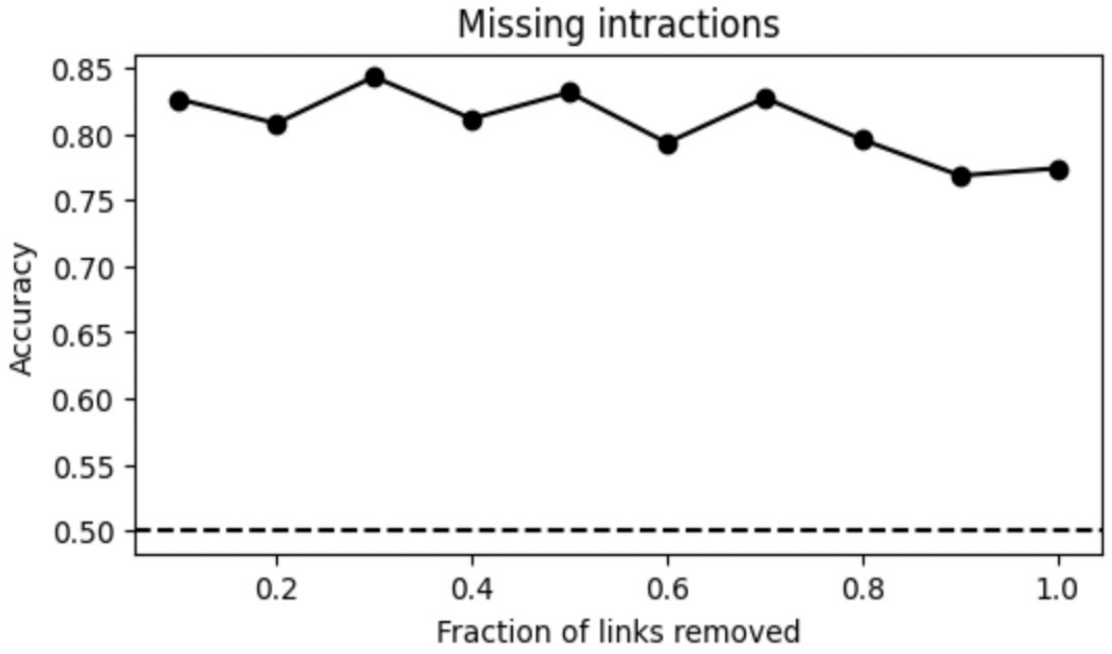


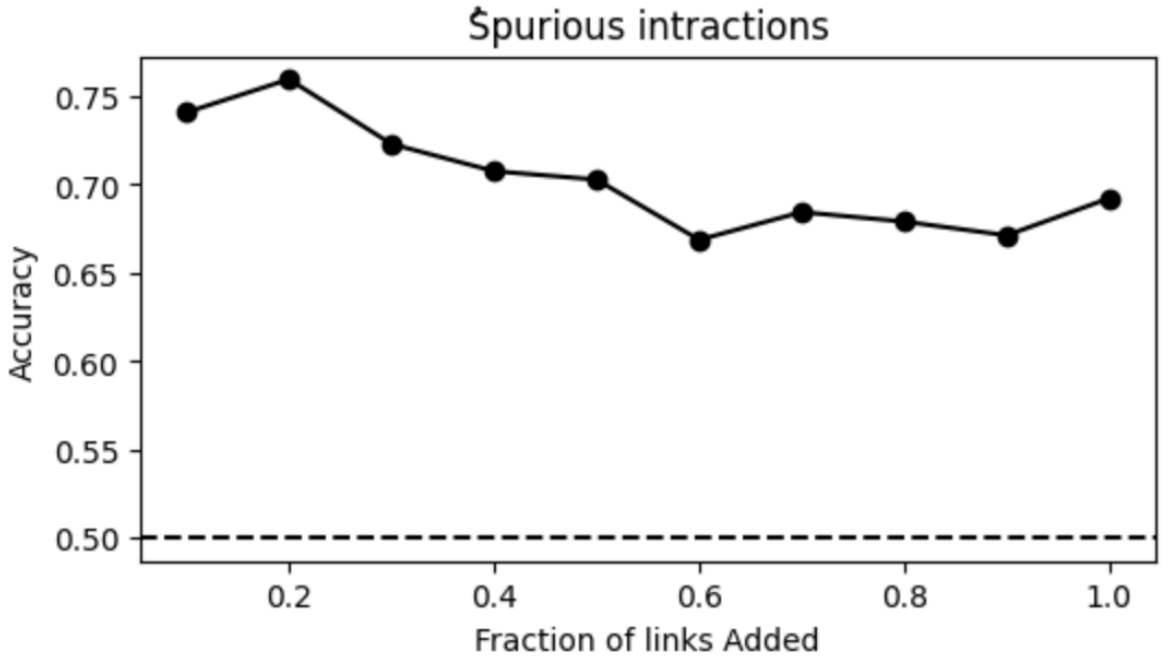




RESULTS

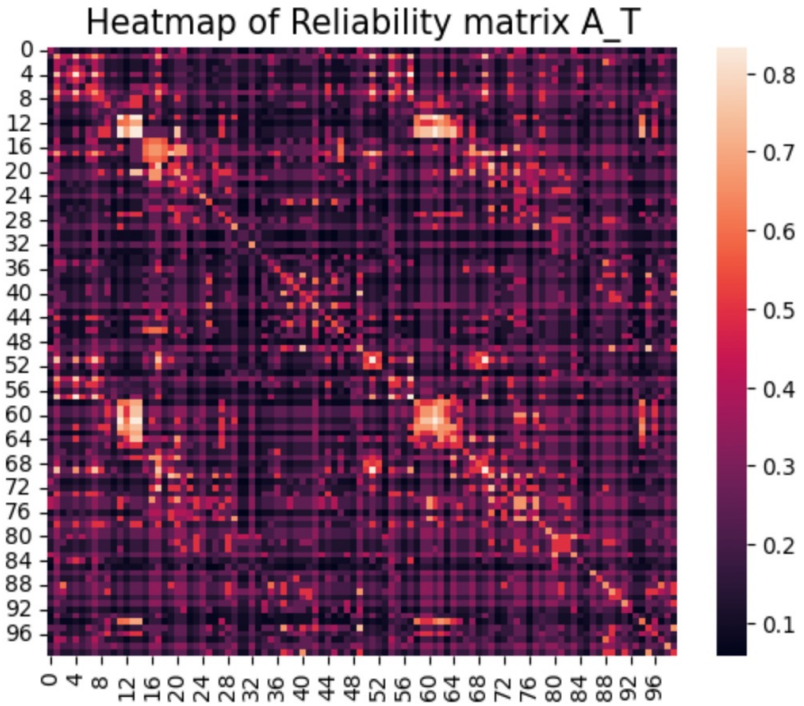
- Missing intractions :
  - Removing some links from the True network in different ratio (Zero-Fake , Zero-Real)
  - False negetive has higher ranking than True negetive
- Spurious intractions :
  - Adding some links from the True network in different ratio (One-Fake , One-Real)
  - False positive has a lower ranking than True positive



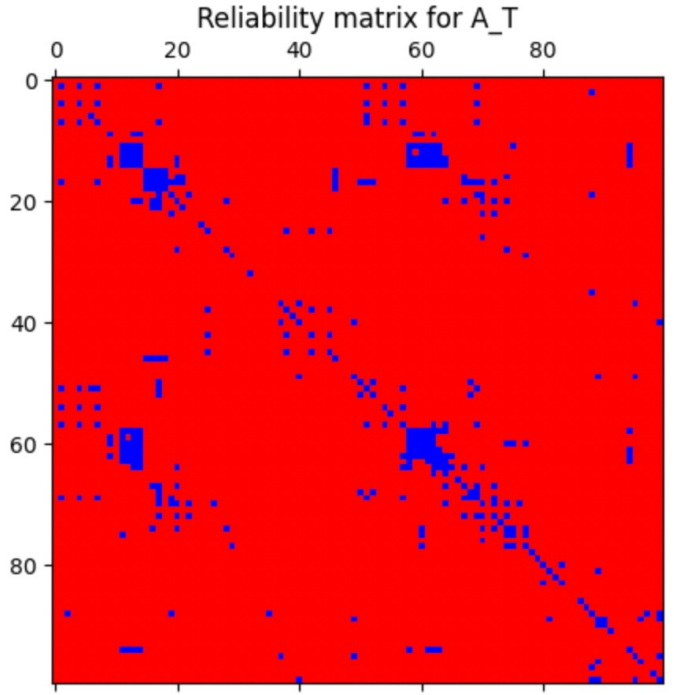


RESULTS

Calculating the Reliability Matrix for True-network

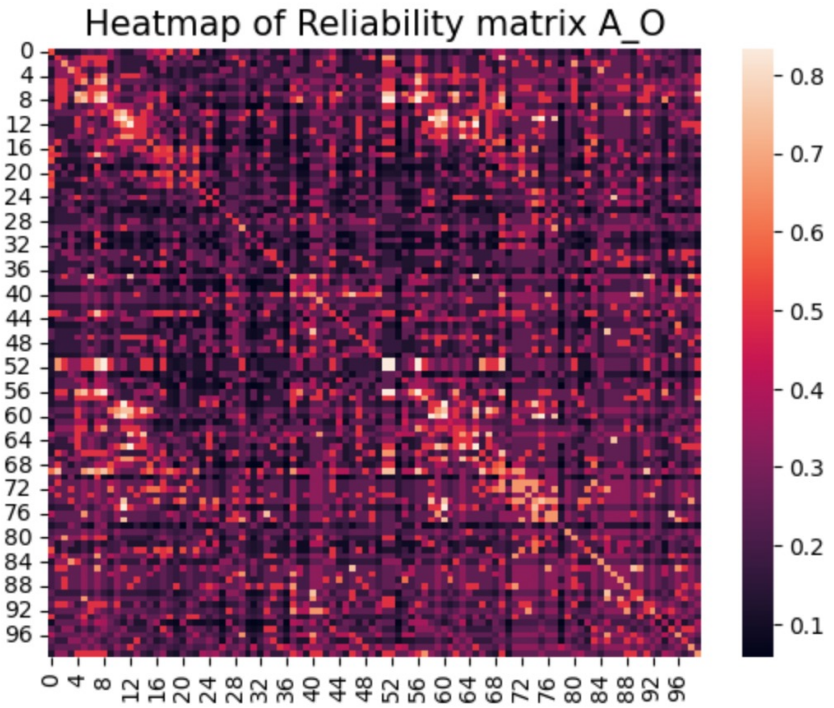



```
A_T[np.where(R_mat_A_T > 0.5)] = 1
A_T[np.where(R_mat_A_T < 0.5)] = 0
```

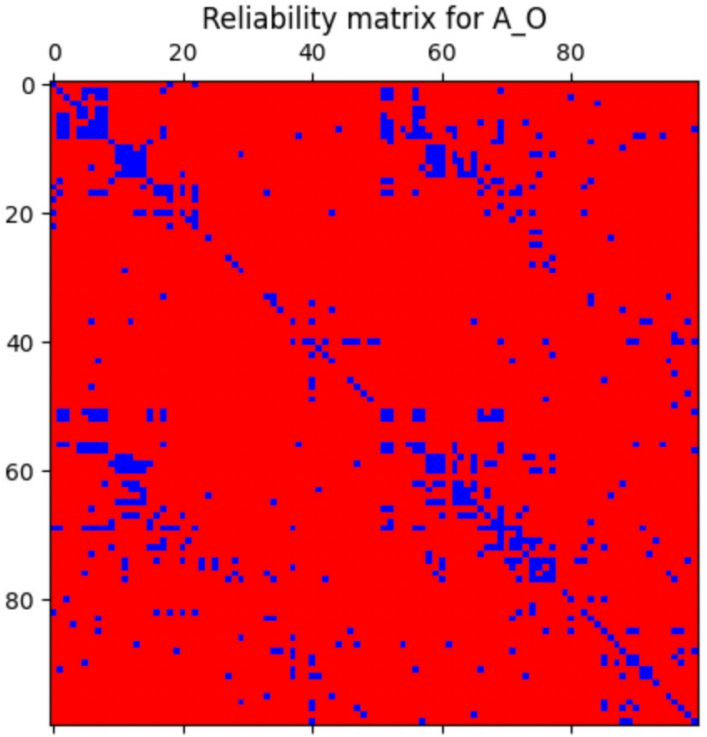


RESULTS

Calculating the Reliability Matrix for Observed-network

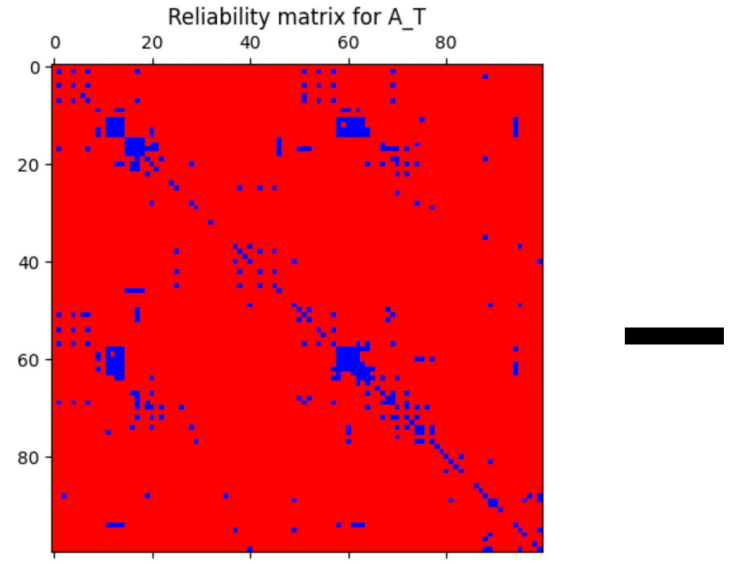


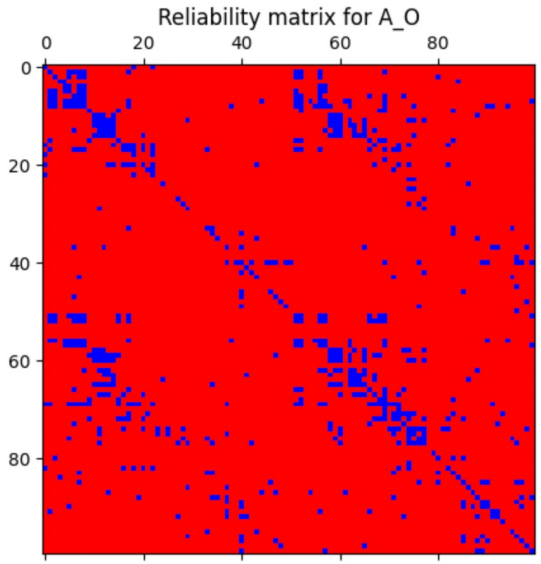

```
A_O[np.where(R_matix > 0.5)] = 1
A_O[np.where(R_matix < 0.5)] = 0
```

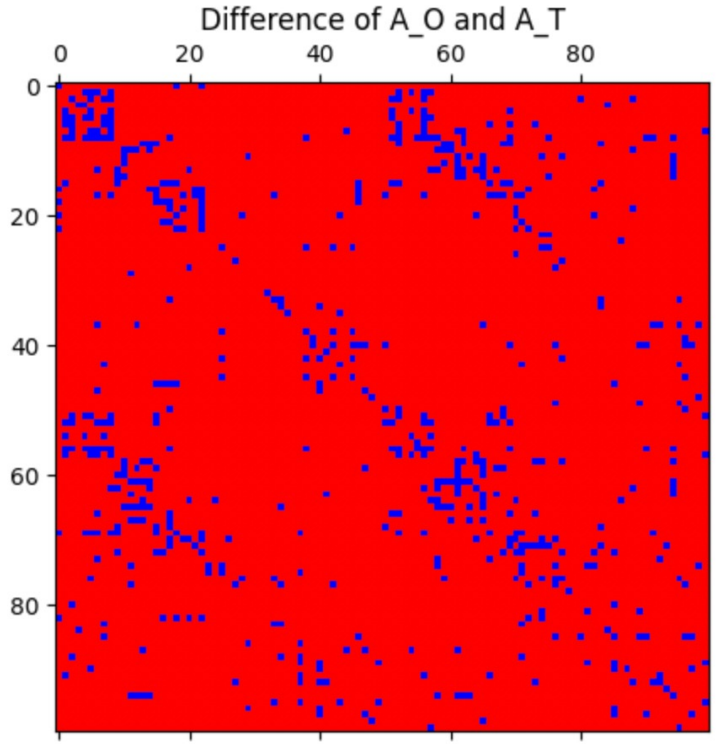


RESULTS

Difference of Observed network and True network







Conclusion

- Investigating the reliability of network
- Building a framework for detecting the missing and spurious connections
- A helpful framework to reconstruct complex networks