9
Network reconstruction using diffusion dynamics

Networks are ubiquitous in the world from natural networks to artificially generated networks [Pagani and Aiello, 2013; Amaral *et al.*, 2000]. Biological, communication, transportation networks, and WWW are few examples of such existing real-world complex networks [Newman, 2010; Girvan and Newman, 2002]. The networks provide a platform to the dynamical processes, such as information diffusion, epidemic spreading, diffusion of resources, to run [Pastor-Satorras and Vespignani, 2001; Wu *et al.*, 2004]. Underlying network structure affects the diffusion processes and helps in controlling them. The power-grid network is a human-made complex network which is an example of the weighted network and the capacity of electricity transportation through a link can be thought as the weight of that link. The distribution of the links and their weights defines the performance of the network in the electricity transportation. A diffusion process is governed by the rules of diffusion and structural properties of the underlying network. Identification of the networked non-linear systems is fundamental to control them. In many cases, exact information regarding the network structure and its nodal dynamics are unknown. To develop a better understanding regarding how to control the networked systems, nodes' interconnection pattern and their properties are required. Diffusion dynamics, for example, information diffusion and epidemic spreading provide time series data which contains the status of the nodes at different time steps. If a node is infected or informed, its status will be 1 otherwise 0. It is a binary data and called as status-time-series (STS) data. Reconstruction of the structure of a network using available status-time-series data of a diffusion process that is seen in the network is an interesting inverse problem and studied in many contexts [Tomovski and Kocarev, 2015; Timme, 2007; Li *et al.*, 2015].

The practical importance of the network reconstruction in many disciplines motivated us to propose a more generalized method for network reconstruction. As a contribution, in this chapter, we propose a simple yet efficient method of network reconstruction using just STS (binary) data. The proposed method does not require any additional information such as the selection of threshold values in compressed sensing theory (CST) based method of network reconstruction. The proposed network reconstruction method is tested on the networks of different structural properties, for example, modular network, random networks, and real-world networks and a high accuracy is achieved in most of the cases. Further, the proposed method is compared with a CST based method of network reconstruction [Shen *et al.*, 2014] using the same STS data and observed superior performance.

In a social network, tracking of information diffusion or rumor spreading with their engaged propagation paths is a tedious task and in some cases, it is almost impossible, for example, viral infection. It is rather easy to keep the record of which node is getting information or infection at what time which is known as the status of the nodes. Using the information of the status of the nodes, the question arises if we can track the whole propagation network or simply network structure in which links and nodes are actively participating? The same problem can be seen as the *detection of missing links* in a network using time series data of a dynamical process attached to that network.

Usually, problems on diffusion dynamics are formulated in the following way: For a

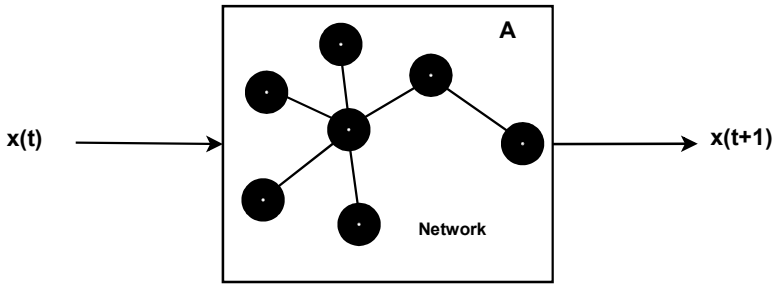


Figure 9.1 : Graphical representation of diffusion dynamics in a network in which the network provides a platform for the diffusion. Output of the diffusion at time step t, $\mathbf{x}(t)$ acts as an input for the next time step.

given topological structure and nodal dynamics what is the collective dynamics of the network? Synchronization, flocking, consensus, cascading failure are some examples of this type of problems [Ren *et al.*, 2005b; Olfati-Saber, 2006; Xi *et al.*, 2010]. Reconstruction of network structure is an example of the inverse problem which has importance in many practical applications in many disciplines like genetic engineering, biological systems, due to the existence of a networked structure seen in dynamical systems. Sometimes it is not easy to collect information regarding the structure of a network, for example, brain network, gene regulatory network. In such cases, reverse engineering can help to get information about underlying network structure.

In this chapter, we develop a method for reverse engineering of network reconstruction based on diffusion dynamics and matrix analysis, only using binary time series data. SIS model [Pastor-Satorras and Vespignani, 2001] is considered as an example to examine the procedure. We start from the introduction of SIS diffusion model at node level given by $\dot{x}_i(t) = \gamma \sum_j a_{ij} x_j(t) - \delta x_i(t)$. In this model, a node can be either in the susceptible state (S) or infected state (I). $x_i(t)$ is the probability of the node i being infected or informed at time t. The collective dynamics of SIS model is given by [Newman, 2010],

$$\dot{\mathbf{x}}(t) = \gamma A \mathbf{x}(t) - \delta I \mathbf{x}(t). \tag{9.1}$$

where a_{ij} is ij^{th} element of the connection (adjacency) matrix, A, of the considered network. γ is diffusion probability of an edge and δ is recovery probability of an infected node that directly goes to susceptible class. $\mathbf{x}^{\mathrm{T}}(t) = [x_i(t), \, , ..., \, , \, x_n(t)]$ is the state vector. For a given time series, $x(t)$ and $\dot{x}(t)$ are known. Eq. (9.1) is linear in coefficient of matrix A. From sufficient amount of data, we can obtain matrix A using the matrix generated by deterministic vectors $\mathbf{x}(t)$ and $\dot{\mathbf{x}}(t)$ that is explained further in this chapter. A graphical representation of discrete time diffusion dynamics is as shown in Fig. 9.1 where $\mathbf{x}(t)$ acts as an input for the next time step, $t+1$. Discrete time diffusion dynamics is adopted to calculate the vector $\dot{\mathbf{x}}(t)$ by considering the approximation $\dot{\mathbf{x}}(t) = \Delta\mathbf{x} = \mathbf{x}(t+1) - \mathbf{x}(t)$. Vectors $\mathbf{x}(t)$ and $\dot{\mathbf{x}}(t)$ are not binary vectors but we have only binary data. It is assumed that at a given time step t binary status vector $\mathbf{y}(t)$ is the resultant of polarization of vector $\mathbf{x}(t)$ in which higher values of $x_i(t)$ are replaced with 1 and lower ones are replaced with 0 using some threshold value which is unknown to us. The benefit of this assumption is that we do not need the information of threshold value as we have already polarized vector. We can also assume that $\mathbf{y}(t)$ is the resulting vector of linear or non-linear transformation of vector $\mathbf{x}(t)$. We define a matrix analysis based simple procedure to reconstruct the connection matrix from the given binary time series data. The procedure is successfully tested in many real-world and computer-generated benchmark networks. The proposed network reconstruction method has higher accuracy and efficiency as compared to the other methods that are considered in this chapter.

Rest of the chapter is arranged as follows: The proposed method of network reconstruction from available status-time-series (STS) data is explained in Section 9.1. Section 9.2 provides the performance analysis. Discussion and salient features of the proposed network reconstruction method are discussed in Section 9.3. Finally, the chapter is concluded in Section 9.4.

9.1 METHODOLOGY OF NETWORK RECONSTRUCTION FROM THE GIVEN STS DATA

Consider a simple SIS model of diffusion in which a node can be in two states, susceptible (S) or infected (I). Let us assume that a node j is in infected state with probability $x_j(t)$ at time t and passes the infection (information or signal) to its neighbouring nodes with probability γ or itself goes to the susceptible state with probability δ. The probability of a node i, being infected at time $t+1$ will be

$$\dot{x}_i(t) = \gamma \sum_j a_{ij} x_j(t) - \delta x_i(t),$$

where a_{ij} is the ij^{th} entry of the connection (adjacency) matrix, A, [Newman, 2010]. The collective diffusion dynamics in the network is given by,

$$\dot{\mathbf{x}}(t) = \gamma A \mathbf{x}(t) - \delta \mathbf{x}(t). \tag{9.2}$$

where A is the connection matrix of the network and vector $\mathbf{x}(t)$ is the probabilities of the nodes of being infected at time t [Newman, 2010]. Consider discrete values of time $t = 0, 1, 2, ..., \tau$, that produce a system of linear equations given by

$$\begin{aligned}
&\dot{\mathbf{x}}(0) = (\gamma A - \delta \mathrm{I})\,\mathbf{x}(0),\\
&\dot{\mathbf{x}}(1) = (\gamma A - \delta \mathrm{I})\,\mathbf{x}(1),\\
&\quad ..\\
&\quad ..\\
&\dot{\mathbf{x}}(\tau) = (\gamma A - \delta \mathrm{I})\,\mathbf{x}(\tau).
\end{aligned}$$

Above given equations can be combined as

$$[\dot{\mathbf{x}}(0)\ \dot{\mathbf{x}}(1)\ ...\ \dot{\mathbf{x}}(\tau)] = (\gamma A - \delta \mathrm{I})\,[\mathbf{x}(0)\ \mathbf{x}(1)\ ...\ \mathbf{x}(\tau)]\,.$$

Let $\mathrm{X} = [\mathbf{x}(0)\ \mathbf{x}(1)\ ...\ \mathbf{x}(\tau)]$ and $\Delta\mathrm{X} = [\dot{\mathbf{x}}(0)\ \dot{\mathbf{x}}(1)\ ...\ \dot{\mathbf{x}}(\tau)]$ be the matrices of size $n \times \tau$. We get a simple equation

$$\Delta\mathrm{X} = (\gamma A - \delta \mathrm{I})\,\mathrm{X}, \tag{9.3}$$

in which X is not a square matrix, so we cannot get the inverse of X directly. To solve this problem, we multiply X^{T} both sides of Eq. (9.3) which produces

$$\Delta\mathrm{X}\mathrm{X}^{\mathrm{T}} = (\gamma A - \delta \mathrm{I})\,\mathrm{X}\mathrm{X}^{\mathrm{T}}.$$

$\mathrm{X}\mathrm{X}^{\mathrm{T}}$ is an $n \times n$ matrix. Next, time t should be divided in τ intervals such that $\dot{\mathbf{x}}(t) \approx \mathbf{x}(t+1) - \mathbf{x}(t) \neq 0,\ \forall\, t$, to avoid redundant calculations and unnecessary engagement of computing and storage resources. Ideally, τ should be selected in such a way that

$$\min_\tau rank(\mathrm{X}_{n\times\tau}) = n,$$

which reduces the unnecessary redundancy and computational complexity of the matrix multiplications. Consider,

$$\gamma A = \left(\Delta \mathrm{X}\mathrm{X}^{\mathrm{T}}\right)\left(\mathrm{X}\mathrm{X}^{\mathrm{T}}\right)^{-1} + \delta \mathrm{I}. \tag{9.4}$$

From Eq. (9.4), it is required that $\mathrm{X}\mathrm{X}^{\mathrm{T}}$ should be invertible matrix. Let τ be sufficiently large such that X has rank n then using Sylvester's rank inequality [Mirsky, 2012]

$$rank(\mathrm{X}\mathrm{X}^{\mathrm{T}}) = rank(\mathrm{X}).$$

$\mathrm{X}\mathrm{X}^{\mathrm{T}}$ is a full rank matrix which confirms the existence of $\left(\mathrm{X}\mathrm{X}^{\mathrm{T}}\right)^{-1}$.

In Eq. (9.4), A is a matrix of zeros and ones while $\left(\Delta \mathrm{X}\mathrm{X}^{\mathrm{T}}\right)\left(\mathrm{X}\mathrm{X}^{\mathrm{T}}\right)^{-1}$ can have real numbers. In this situation, a transfer function, which simply applies a threshold on the values of the considered matrix (the right side of Eq. (9.4)), is used to convert values to zeros and ones to get adjacency matrix A. Let suppose, the transfer function is given by

$$A = F\left(\left(\Delta \mathrm{X}\mathrm{X}^{\mathrm{T}}\right)\left(\mathrm{X}\mathrm{X}^{\mathrm{T}}\right)^{-1}\right). \tag{9.5}$$

Similarly, status-time-series (STS) data, Y, can be assumed as the transformation of X and given by,

$$\mathrm{Y} = \Phi(\mathrm{X}). \tag{9.6}$$

From Eqs. (9.5) and (9.6), $\exists$ a Ψ such that

$$A = \Psi(\mathrm{Y}), \tag{9.7}$$

which would not break the consistency of the theory. In this chapter, we attempt to find a simple relation between A and Y. In other words, we try to find out an approximation of Ψ using the structure of Eq. (9.4).

In the case of the real-world diffusion process, we have status-time-series (STS) data of the nodes in which status of the nodes that they are infected or susceptible are given at the particular time. It is represented by numbers 0s and 1s.

We adopted the structure of Eq. (9.4) to define a method to reconstruct the connection matrix of a network using times series data of the status of the nodes during the diffusion process. Matrix M is considered as an approximation of Ψ. The procedure of network reconstruction is discussed in Algorithm 1.

The proposed algorithm depends on STS data (Y) and the number of edges (m). In Section 9.2.4, we discuss the retrieval of the number of edges from the STS data only. The proposed methodology of inferring network topology from STS data is simply based on observations (diffusion data). The algorithm does not require prior knowledge of nodal dynamics or diffusion

Algorithm 1 Network Reconstruction

Input: STS data (Y), outcome of the diffusion dynamics on a network G.
OutPut: Connection matrix C.

1: **procedure** Reconstruction
2: Consider matrix $M = \left(\Delta \mathrm{Y}\mathrm{Y}^{\mathrm{T}}\right)\left(\mathrm{Y}\mathrm{Y}^{\mathrm{T}}\right)^{-1}$. Size of the matrix Y is $n \times \tau$ which contains time series data of the status of the nodes and column of the matrix ΔY represents the changes in the status of the nodes in consecutive time steps.
3: $M \leftarrow \dfrac{M + M^{\mathrm{T}}}{2}$, to generate a symmetric matrix.
4: In the decreasing order of the entries of the matrix M, select first $2m$ entries from the matrix M. Replace these entries with 1 and rest of the entries with 0. Newly produced matrix of 0s and 1s is called reconstructed connection matrix C of the given network G.
5: **end procedure**

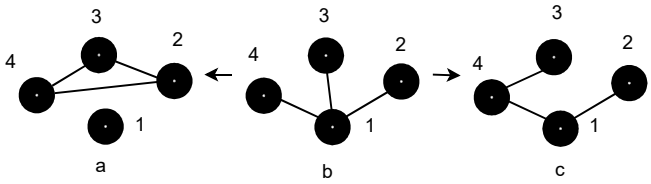


Figure 9.2 : (a) Example of wrongly reconstructed network. (b) Considered network (original). (c) Example of reconstructed network which is partially correct.

parameter as required in other considered methods in the literature [Timme, 2007; Shandilya and Timme, 2011; Aniszewska and Rybaczuk, 2008; Zhou and Lu, 2007; Comellas and Diaz-Lopez, 2008].

Another aspect of the network reconstruction framework is the evaluation of the accuracy of Algorithm 1. True Positive Rate (TPR) and False Positive Rate (FPR) are considered as measures or metrics to evaluate the accuracy of the proposed algorithm. TPR is defined as the fraction of links correctly identified and FPR is the fraction of links wrongly identified which do not exist in real. We define another measure to check the accuracy of the reconstruction of the network which inherits the flavor of TPR and FPR both, simultaneously and given by

$$\Delta E = 1 - \frac{\mathbf{1}^{\mathrm{T}} abs(A - C)\mathbf{1}}{\mathbf{1}^{\mathrm{T}} A \mathbf{1}}, \tag{9.8}$$

where $abs(A - C) = (|a_{ij} - c_{ij}|)$, c_{ij} is ij^{th} entry of the matrix C. ΔE has maximum value 1 and minimum value -1. If the reconstruction is done accurately then ΔE will be 1 and in case of totally wrong reconstruction, it will be -1. $\mathbf{1}^{\mathrm{T}} abs(A - C)\mathbf{1}$ counts the mismatches in the constructed network and original network. Accurate reconstruction of the network eliminates the second term in Eq. (9.8) due to no mismatch of reconstructed links.

The effectiveness of the measure ΔE is explained by an example in Fig. 9.2. An example network of 4 nodes and 3 edges is considered. Adjacency matrix of original network is given by

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

and reconstructed networks in Fig. 9.2 have connection (adjacency) matrices

$$C_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \text{ (Fig. 9.2a) and } C_2 = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \text{ (Fig. 9.2c),}$$

respectively.

$$A - C_1 = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & -1 & -1 \\ 1 & -1 & 0 & -1 \\ 1 & -1 & -1 & 0 \end{bmatrix},$$

$$abs(A - C_1) = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix},$$

$$\Delta E = 1 - \frac{\mathbf{1}^{\mathrm{T}} abs(A - C_1)\mathbf{1}}{\mathbf{1}^{\mathrm{T}} A \mathbf{1}} = 1 - 2 = -1.$$

Similarly, for the reconstructed matrix C_2,

$$\Delta E = 1 - \frac{\mathbf{1}^{\mathrm{T}} abs(A - C_2)\mathbf{1}}{\mathbf{1}^{\mathrm{T}} A \mathbf{1}} = 1 - 2/3 = 1/3.$$

In the next section, we provide the performance analysis of the proposed method of network reconstruction.

9.2 PERFORMANCE ANALYSIS

We generated networks using the models LFR (Lancichinetti-Fortunato-Radicchi) [Lancichinetti *et al.*, 2008], GN (Girvan-Newman) [Girvan and Newman, 2002], ER (Erdős-Rényi) [Erdös and Rényi, 1960], BA (Barabási-Albert) [Barabási *et al.*, 2000] and WS (Watts and Strogatz) [Watts and Strogatz, 1998a] which provide the benchmark for testing of different type of algorithms and diffusion dynamics such as community detection, graph decomposition, epidemics and information diffusion. SIS model is applied on the computer generated networks of the models LFR, GN, ER, BA, and WS and other considered real-world networks given in Table 9.1. The process of diffusion is simulated under the SIS model and STS data of diffusion is collected for each of the above mentioned network and the proposed network reconstruction procedure, Algorithm 1, is applied.

The graphical representation of the framework that is used to validated the proposed network reconstruction method, is given in Fig. 9.3. The proposed framework has three parts dedicated to three separate processes given as

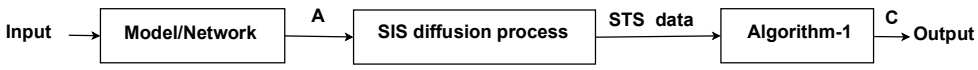


Figure 9.3 : Graphical representation of the framework that is used to validated the proposed network reconstruction method. Accuracy ΔE is calculated which counts the mismatches in the entries of adjacency matrix A and reconstructed matrix C.

1. Consideration of a network, real or computer generated, under a considered model.
2. Diffusion is applied under SIS diffusion model to get STS data.
3. Network reconstruction is performed using Algorithm 1 and accuracy ΔE is calculated.

A detailed description of the network generation using the models is given in Section **??**. Details of the implementation of diffusion under SIS and proposed method of network reconstruction is given in Section 9.2.2.

9.2.1 Network models used for simulation of reconstruction process

Network reconstruction method is applied to different networks in which some are real world and some are computer generated networks. We consider GN [Girvan and Newman, 2002], ER [Erdös and Rényi, 1960], WS [Watts and Strogatz, 1998a], BA [Barabási *et al.*, 2000] and LFR [Lancichinetti *et al.*, 2008] benchmark model networks.

GN benchmark networks are regular random networks with community structure. The size of GN networks is 128 nodes in which each node has the same degree. GN model network consists of 4 communities of the same size, 32 nodes each. Community strength depends on the ratio of internal to the external degree of nodes. We used a GN benchmark network of average degree 16 in which each node has 6 connections inside the community and 10 connections outside the community.

ER and WS, both follow binomial degree distribution. ER model network is produced by applying Bernoulli process over each pair of nodes. Consider a pair of nodes and linked them with probability $p = 0.1$. In this way, we generated a connected ER model network of 100 nodes. WS networks are the resultant of rewiring of regular networks with probability $p > 0$ which exhibit small world phenomena having higher clustering as real-world networks have [Newman, 2010]. First, a regular network of desired average degree is considered then rewiring process in applied. Each edge is rewired with probability $p = 0.001$. In the rewiring process, an existing edge disappears and a new edge appears randomly. A network of average degree 8 is considered under this model.

BA model and LFR has power-law degree distribution. BA model is a growing random network model that is based on preferential attachment scheme in which, at each time step, a node appears and gets connected to a pre-existing node i with probability p_i^j given by

$$p_i^j = \frac{k_i}{\sum_l k_l},$$

where k_i is the degree of node i. We considered a network of average degree 4 that is generated by BA model. LFR network is generated using the procedure given in [Lancichinetti *et al.*, 2008] by setting the parameters: number of nodes 100, average degree 5, maximum degree 20, exponent for

Table 9.1 : Accuracy of the reconstruction of given networks. n and E are the number of nodes and edges respectively in a network. ΔE is the accuracy measure of the reconstruction. The provided values are averaged over 100 runs of reconstruction algorithm for each network of given size. The parameters are $\gamma = 0.1$ and $\delta = 0.03$. The values of the accuracy measures (Δ, TPR and FPR) in 5^{th}, 6^{th} and 7^{th} columns are corresponding to our proposed method.

	Models/ Networks	n	m	ΔE	TPR	FPR	ΔE (CST)
Network models	BA model	500	986	1	1	0	0.984
	ER model	100	473	0.9789	0.9895	0.0105	0.963
	GN	128	1024	1	1	0	0.956
	LFR	100	203	1	1	0	1
	WSM	500	2000	1	1	0	1
Real-world networks	Col	145	346	1	1	0	0.974
	Dolphins	62	159	1	1	0	0.968
	Celegans	453	2025	1	1	0	0.993
	Football	115	613	1	1	0	0.971
	Web	180	228	0.9956	0.9978	0.0022	0.959
	Lesmis	77	254	1	1	0	0.989
	Karate	32	78	1	1	0	0.9744

the degree distribution 2, exponent for the community size distribution 1, mixing parameter 0.1, number of overlapping nodes 0, number of memberships of the overlapping nodes 0 and range of community size is taken as $[5,\ 20]$. The details of the data are given in Table 9.1.

9.2.2 STS data generation and network reconstruction

Susceptible-Infected-Susceptible (SIS) diffusion model is adopted to generate time series status data of diffusion over the considered networks. Initially, all nodes are in susceptible (S) state or class. A node is randomly selected from the network as an initial spreader (source node). During diffusion phenomena, a node, which is in the infected class of nodes (I), can infect its neighbor with probability γ or it can be transferred to susceptible state with probability δ. Simulation is done by setting $\gamma = 0.1$ and $\delta = 0.03$. At each time step t, all infected nodes either try to transfer their infection to their susceptible neighboring nodes with probability $\gamma = 0.1$ or shift to susceptible state with probability $\delta = 0.03$. After completion of each step, we collect the status of the nodes, either they are infected ($y_i(t) = 1$) or susceptible ($y_i(t) = 0$). We get $\mathbf{y}(t)$ for each time step which collectively provide the matrix Y. After that we generate ΔY with the help of matrix Y.

$$\Delta \mathrm{Y} = \mathrm{Y}(:, 2 : \tau + 1) - \mathrm{Y}(:, 1 : \tau).$$

Now we apply the proposed network reconstruction method (as stated in Algorithm 1) over the generated STS data. We get the matrix M of real values. $2m$ highest values are selected to convert them into ones and rest of the values set to zeros. The resulting matrix is called as reconstructed connection matrix C. The performance of the network reconstruction method is evaluated using well-accepted measures TPR and FPR, and our defined measure, ΔE, given in Eq. (9.8). Results are provided in Table 9.1. 9 out of 11 considered networks are reconstructed with 100% accuracy ($\Delta E = 1$, TPR$= 1$ and FPR$= 0$). The other two networks, Web, and ER are reconstructed with high accuracy having the values of accuracy measure $\Delta E = 0.9956$ and 0.9789, respectively.

The value of γ and δ can vary without affecting the accuracy of the proposed reconstruction

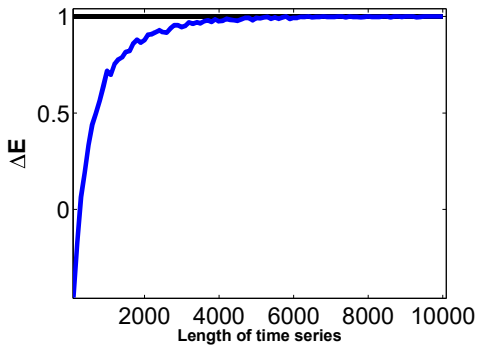


Figure 9.4 : Improvement in accuracy of network reconstruction from time series data of diffusion. SIS diffusion model is applied in KARATE network to produce STS data. Horizontal axis represents the length of status-time-series (STS) data and vertical axis corresponds to accuracy (ΔE) of network reconstruction.

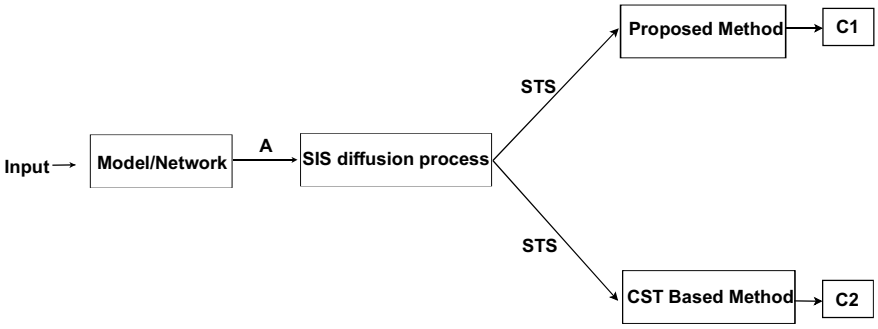


Figure 9.5 : Comparison of two methods of network reconstruction. One is our proposed method ($C1$) and second is CST based network reconstruction ($C2$).

method. We should select the values of parameters γ and δ such that STS data that is generated has sufficient independent vectors $\mathbf{y}(t)$. The reconstruction of the considered networks are tested for other values of the parameters and results are similar to given in Table 9.1.

The accuracy of the method depends on the amount of available data. From the given sufficient length of time series data of diffusion, we can able to reconstruct the respective network without any error. For an example, a network of KARATE club member [Newman, 2006b] is considered to evaluate the accuracy of the reconstruction of the network, with respect to available status-time-series data, which is shown in Fig. 9.4. From Fig. 9.4, we can clearly observe that the accuracy (ΔE) of the proposed network reconstruction method increases with the length of time series data.

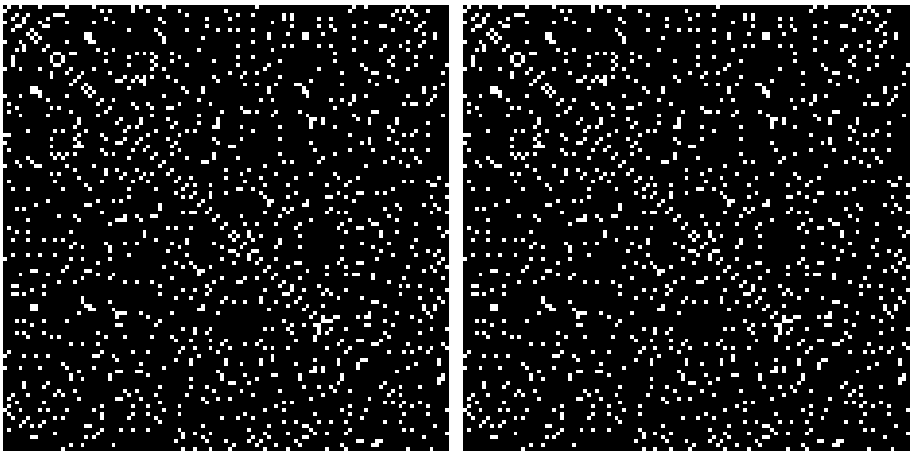
9.2.3 Efficiency and accuracy of the proposed algorithm

In this section, we perform a comparative analysis of the proposed method of network reconstruction and an existing CST based method for the same. STS data is used as the input in both the methods (our proposed one and CST based method).

The success rate of the complete reconstruction of networks using the proposed method is almost 82% (9 out of 11 networks) while the considered CST based method has success rate of 16.6% (2 out of 11 networks), refer Table 9.1. The success rate of complete reconstruction of networks using CST based method reported by the authors was 33% (4 out of 12) [Shen *et al.*, 2014]. Same STS data is used in both methods (our proposed one and the considered CST based method) of network reconstruction. CST based method of network reconstruction requires two thresholds values to calculate the probability distribution of infection and selection of independent input vectors. For more detail refer [Shen *et al.*, 2014]. Our proposed method does not require any prior information regarding the selection of threshold values. The accuracy of the proposed method is better as compared to that of the CST based network reconstruction method. The accuracy of the CST based method depends on the selection of threshold values. A graphical representation of the performance comparison of the two methods of network reconstructions is given in Fig. 9.5. In Fig. 9.5, $C1$ and $C2$ are reproduced adjacency matrices which are obtained by the proposed and the considered CST based network reconstruction methods, respectively. $C1$ and $C2$ are compared on the basis of metric ΔE in Table 9.1. The strength of the proposed method of network reconstruction can be followed by comparing the accuracy values given in fifth (our proposed method) and eighth (CST based method) column of Table 9.1.

An example of reconstructed network

Here we demonstrate the accuracy of the proposed method of network reconstruction via an example of the real-world network. In Fig. 9.6(a), the structure of the original adjacency matrix of the football network is shown and Fig. 9.6(b) is corresponding to reconstructed adjacency matrix also known as connection matrix C. Football network has 115 nodes and 613 edges. This is a social network among the football players who participated in a football tournament. First, diffusion processes imposed in the considered network then the connection matrix C is obtained using Algorithm 1. Considered adjacency matrix and reconstructed connection matrix have the exact overlapping of zeros (black dots) and ones (white dots) in Fig. 9.6. TPR is 1 and FPR is 0 which is corresponding to 100% accuracy of the network reconstruction.



(a) (b)

Figure 9.6 : (a) Adjacency matrix, A, of football network. (b) Reconstructed Adjacency matrix, C.

Previously, we discussed that Algorithm 1 requires STS data Y and number of edges m in advance. Now the question is that can we get the value of m from the available STS data itself?

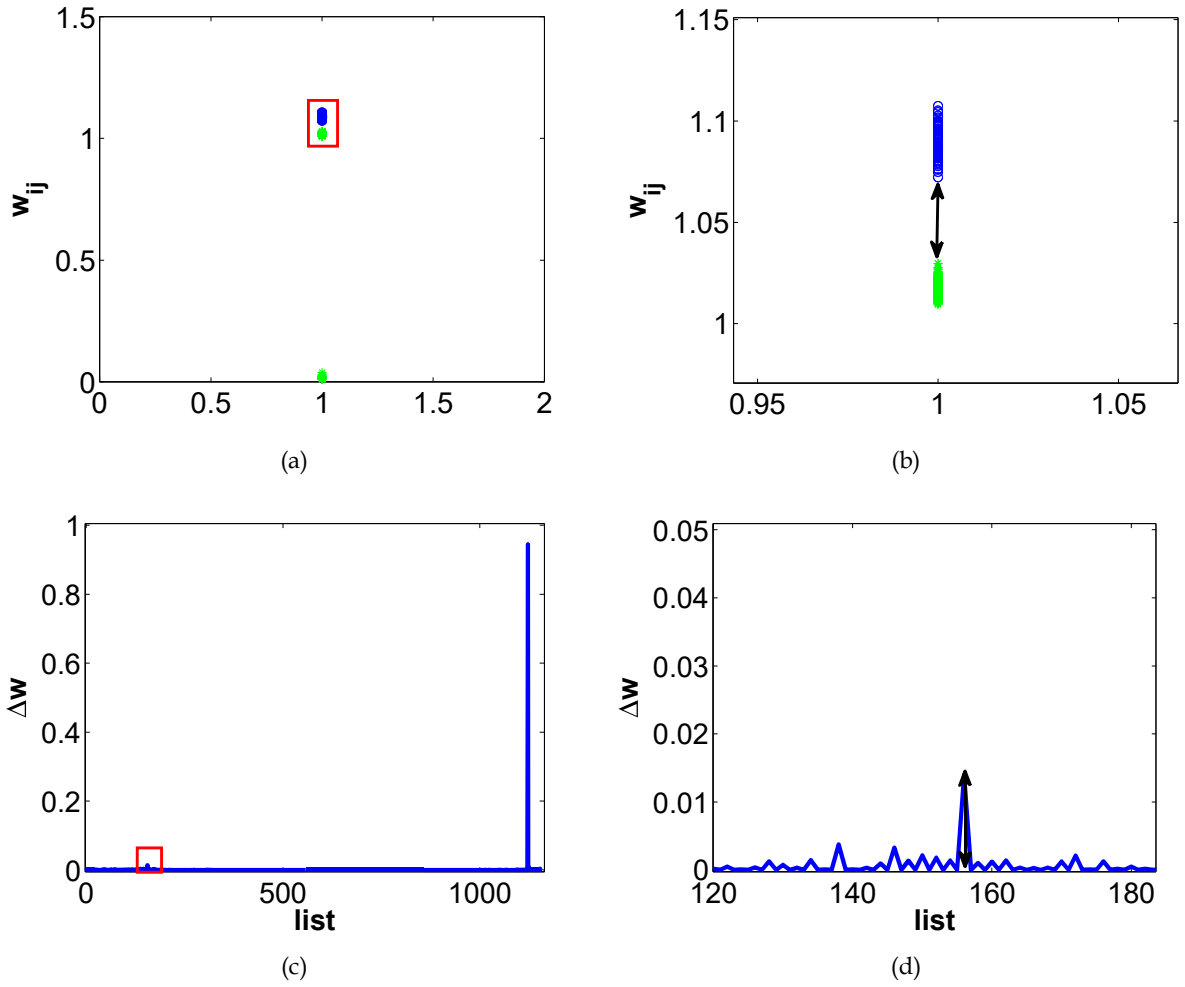


Figure 9.7 : Calculation of the threshold value to generate connection matrix C from the matrix M correspond to KARATE club network of $n = 34$ nodes and $m = 78$ edges. (a) W_{ij} Weights of matrix M. (b) Magnified view of plot inside the red-rectangle in 9.7(a) (Left). (c) Difference of consecutive weights (ΔW) of matrix M when these are sorted. Second highest peak is inside the red-rectangular. (d) Magnified view of plot inside the red-rectangle in 9.7(c) (Left). The height of the arrow represents the difference shown in 9.7(b) (Top).

9.2.4 How to find out the value of m (number of edges)?

In the proposed network reconstruction method, we assume that the number of edges in the network is known in advance. Now the question is that can we reconstruct a network successfully without knowing the number of edges that are present in the network. We plotted the values M_{ij} of the matrix M in Fig. 9.7(a). Fig. 9.7(b) is showing the magnified view of the rectangular area in Fig. 9.7(a). The dots outside the rectangular box in Fig. 9.7(a) are well separated from the rest of the plotted dots (points). These points (lying outside the box) correspond to the strength of the communicability of the pairs of the nodes in the context of diffusion under SIS dynamics. Inside the box, there are two groups of points which are clearly visible in the magnified view shown in Fig. 9.7(b). The upper group of points (in blue color) are representing the edges in the original network and the lower bunch of dots are corresponding to node pairs which have the tendency of the strong indirect relation of communicability in the context of diffusion. A clear separation line between the two groups inside the rectangular box in Fig. 9.7(b) can be drawn to identify the number of edges in the network.

In another observation, we generated a list of length n^2 which has entries of the matrix M of size $n \times n$. Then the difference in the consecutive entries (ΔW) of the sorted list is calculated and plotted in Fig. 9.7(c). Fig. 9.7(d) is the magnified view of the rectangular area in Fig. 9.7(c). There are two noticeable peaks in the plot of Fig. 9.7(c). The position of the first peak (smaller of two peaks) is corresponding to the double of the number of edges, $2m$, in the network without knowing it in advance. Second highest peak (inside the rectangular box in Fig. 9.7(c)) can be used as an identifier to calculate the edges in the network reconstruction. Hence, the information regarding the number of edges, m, is not required in advance. The proposed method of network reconstruction is able to work on STS data only. We applied the above process to find out the value of m for each of the networks considered (as stated in Table 9.1) and observed the exact matching of the number of edges derived from this process with that of the original network.

9.2.5 Prediction of ordering of edges in a weighted network using STS data

Networks are heterogeneous in nature and can have different communication properties of links which can be utilized during communication. For example, let G be a communication network and A is its connection matrix. Edges of the network have different communication or data transfer capacity (weights) which is denoted by w_{ij} and unknown. If we have status-time-series data, how can we retrieve more information? In this section, we are concerned about the ordering of edges according to their weights. The aforementioned method of network reconstruction is applied using status-time-series (STS) data and obtained matrix M. Entries from M, corresponding to non-zero entries in A, are considered as we know the connections of the network. Correlation between original weights and extracted weights from the matrix M is calculated. It is highly positive which suggests that we can successfully order the edges according to their communication importance. This information can be utilized in data communication to prioritize the links while sending the information or data or routing on the Internet.

We consider the real world network of KARATE club members. Weights, which are in the range of $[0\ 1/8]$, are randomly distributed to its edges that work as transmission probability (γ) in diffusion. Each edge has a different value of γ. $0.7(1-\gamma)$ is used as recovery probability, δ, for each infected node. The correlation between recovered weights of edges in the reconstructed network and original weights are highly positive, 0.8439. This indicates that ordering of edges in the reconstructed weighted network can be used to order the edges in the original network of unknown weights. The same process of reconstruction of edge weights is applied to the social network of dolphins. In this case, correlation is 0.8262. These results are averaged over 100 runs of the reconstruction process. The considered length of time series is 10^4.

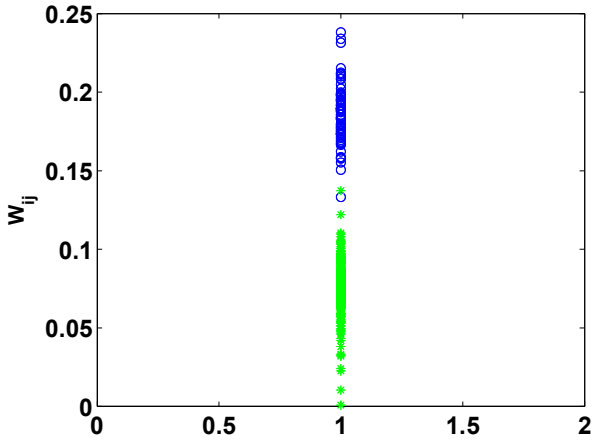


Figure 9.8 : Weights of recovered links using CST based network reconstruction method. There are two group of points in which upper group (blue circles) corresponds to existing links and lower group (green stars) representing the indirectly connected node pairs.

9.3 DISCUSSION

The structure of the matrix M is the resultant of the diffusion dynamics held over the underlying network G, so the distribution of the values M_{ij} signifies the strength of the communicability of each node pair whether they are directly connected or indirectly. Directly connected nodes (existing connections) have highest communication strength which is shown by blue circles in Fig. 9.7(a). Second class or group of points inside the rectangular box in Fig. 9.7(a) (green color stars or lower group of points in Fig. 9.7(b)) represents the indirectly connected pair of nodes of short-range connection (connected by few links). The third group of points at the bottom of the Fig. 9.7(a) are corresponding to long-range indirect connections. The third group of node pairs connected by long-range connections can be focused while improving the connectivity of the network if it is required for an application similar to innovation or information diffusion. There are also multiple applications where the short-range indirect connections are important that are represented by green stars inside the rectangular box in Fig. 9.7(a).

Resulting weights of diffusion dynamics which are the entries of the matrix M also can be utilized to identify the promising pathways of diffusion. If $W_{ij}(=M_{ij})$ has high value then nodes i and j are either directly connected or connected with the highly active pathway of diffusion. To prevent the outbreak condition of the epidemic, we can demolish the highly active pathways and these pathways can be identified by matrix M.

The proposed method is focused on diffusion under SIS (contact based diffusion) model on a network. The method has applicability in many ways: For example, similar matrix analysis based approach can be adopted for other diffusion dynamics. The proposed procedure of network reconstruction is simple in the sense that it needs only to use STS data in Eq. (9.4) which can also be used as a starting point for reconstruction of a weighted network. The method can also be applied to non-STS data where data points can be real numbers instead of binary $(0,1)$. In [Timme, 2007; Shandilya and Timme, 2011; Aniszewska and Rybaczuk, 2008; Zhou and Lu, 2007; Comellas and Diaz-Lopez, 2008; Yu, 2010; Yu and Parlitz, 2011; Li *et al.*, 2015; Ma *et al.*, 2015; Tang *et al.*, 2015], we need to know the nodal dynamics a prior but the proposed one can be applied for more general contact based diffusion dynamics without knowing the exact nodal dynamics. The dependency of

the considered previous methods of network reconstruction on the system-parameter (for example infection rate γ in [Tomovski and Kocarev, 2015]) or nodal dynamics reduces the applicability of the methods. In our proposed method, there is no such constraint. Only STS data is required. The simplicity of the model comes from Eq. (9.4) in which only data is required without any other prior information of the diffusion parameters and nodal dynamics of the nodes.

If we compare the proposed framework with the CST-based method of network reconstruction, we can observe that our proposed method is more informative and able to reveal hidden information. In Fig. 9.8, weights of links obtained by CST based method of network reconstruction are plotted which are separated into two groups. Points of higher weights correspond to existing links and vice versa while in our proposed method, weights are distributed in three groups which are corresponding to directly connected nodes, indirectly connected nodes of strong communicability, and node pairs of weak communicability, see Fig. 9.7(a). In the case of our proposed method of network reconstruction, we also observe that there is a sharp boundary between the three groups of the weights of the node pairs while in CST based method of network reconstruction, the separation line between the weights of direct connections and indirectly connected node pairs is not well defined which makes the identification of the exact number of links present in the network difficult. In the proposed framework of the network reconstruction, second highest peak in the difference of the sorted weights (W_{ij}) of the node pairs is the indicator of the number of links present in the network while there in no such well-defined indicator in CST based network reconstruction method. The selection of two threshold values in CST based methods also makes the procedure less robust [Shen *et al.*, 2014].

9.4 CONCLUSION

Network reconstruction is a problem of great interest which belongs to multiple domains, for example, biology, genetics, social networking, epidemic and wireless sensor network. In this chapter, we presented a simple yet efficient approach to infer the connection matrix from the given binary time series data of a process held over the considered network and also compared with an existing CST based method proposed for the same. From the considered diffusion dynamics under SIS model, a matrix, M, is generated using the status-time-series data that is converted to the connection matrix on the basis of a simple observation made on the entries of the matrix M. Reconstruction procedure has novelty in accuracy and efficiency. The proposed method can be applied in more general contact based diffusion dynamics without prior information of the nodal dynamics. Network topology inference framework can be utilized in link prediction also. Missing links can be identified using the proposed method. In future, the proposed framework can be extended to identify missing nodes as well as links with their meaningful attributes.

…