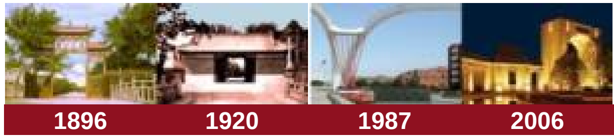


A Broader View of Social Network De-anonymization Problem in Model and Algorithm

5140219173 吴昕宇 大三EE

Outline:

1. Problem formulation

2. Transforming the problem from node matching to edge matching

3. The uniqueness of the optimal edge matching solution

4. The equivalent transformation of the community assignment constraint

5. The convex-concave algorithm to solve this problem

1. Problem formulation

The fundamental model in this paper is as follows. Assume that there is an underlying graph G with n nodes, and the probability of every edge in this graph is p. Based on this graph, we generate two graphs G_1 and G_2, with the edge probability s_1 and s_2 respectively. Note that the edge exists in G_1 or G_2 should satisfy that it exists in G. In G_1, we call it a public network, with the index of every node and the topology known to the attacker, while we call G_2 a target network, with its topology known only and the index of every node unknown(anonymized). The aim of the attacker is to discover the right mapping of indexes of nodes in G_1 and G_2 based on the topological information.

In addition to edges, we also consider communities in two graphs. A community contains several nodes in a graph, which is closer to real social network like Facebook. Therefore the attacker can utilize both community and topology information of both graphs to get the right mapping. Figure 1 shows an instance of the problem.(The figure is cited from Xinzhe Fu's work)

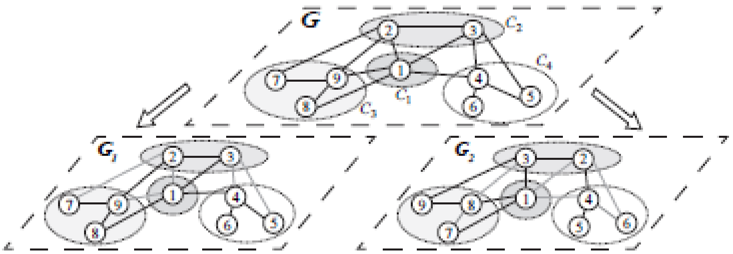


Figure 1: **An example of underlying social network (G), the published network (G_1) and the auxiliary network (G_2) sampled from G. C_1, C_2, C_3, C_4 represent the four communities in the networks. The correct mapping $\pi_0 =$ $\{(1,1),(2,3),(3,2),(4,4),(5,6),(6,5),(7,9),(8,7),(9,8)\}$.**

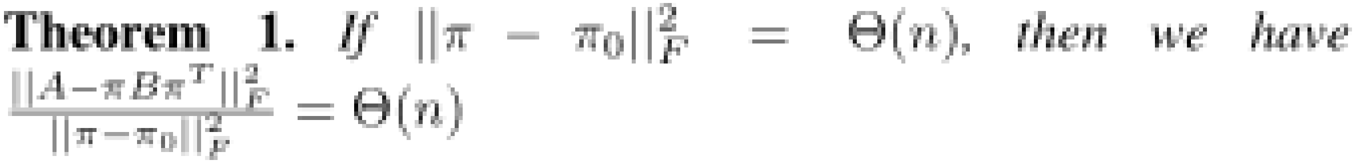
2. Transforming the problem from node matching to edge matching

Incentives:

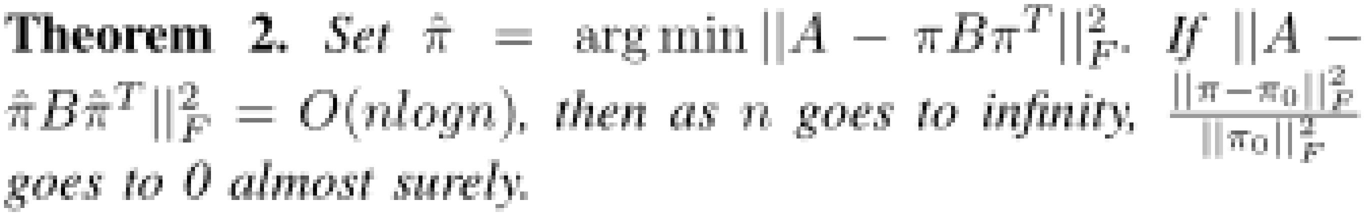
The original problem is a node matching problem. However, no exace information solely based on nodes themselves are known. So we intend to transform it into the edge matching problem, and we discovered that this transformation is reasonable in the aspect of keeping the ratio of the value of these two objective functions $\Theta(n)$

Edge matching problem with communities

$$\begin{aligned} \min \quad & ||A\pi - \pi B||_F^2 + \mu ||\pi M - M||_F^2 \\ s.t. \quad & \pi \mathbf{1} = \mathbf{1}, \quad \pi^T \mathbf{1} = \mathbf{1}. \end{aligned}$$



Theorem 1. *If* $||\pi - \pi_0||_F^2 = \Theta(n)$, *then we have* $\frac{||A-\pi B\pi^T||_F^2}{||\pi-\pi_0||_F^2} = \Theta(n)$



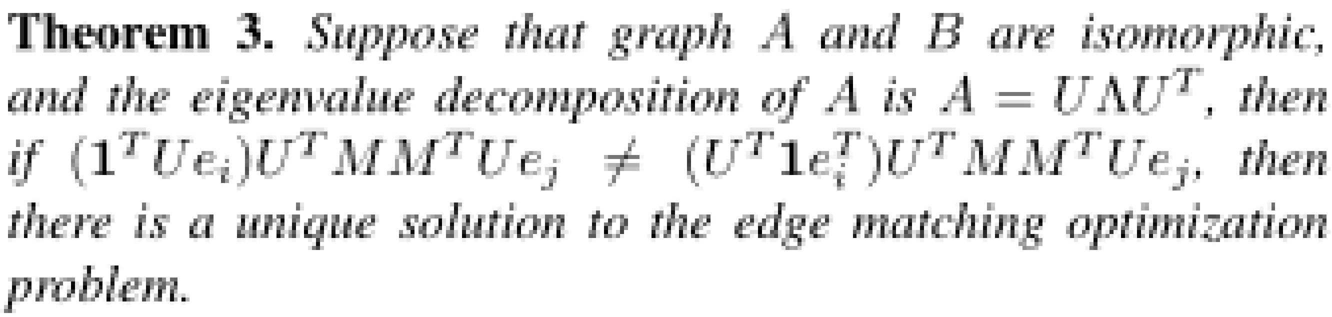
Theorem 2. *Set* $\hat{\pi} = \arg\min ||A - \pi B\pi^T||_F^2$. *If* $||A - \hat{\pi} B\hat{\pi}^T||_F^2 = O(nlogn)$, *then as* n *goes to infinity,* $\frac{||\pi-\pi_0||_F^2}{||\pi_0||_F^2}$ *goes to* 0 *almost surely.*

Theorem 1 discusses the general value of π;
Theorem 2 discusses the optimal solution of the edge matching problem.

3. The uniqueness of the optimal edge matching solution

Incentives:

We should study the uniqueness of the edge matching problem, or else if there are multiple optimal solutions in edge matching problem, it will be hard to determine which is better in node matching problem.

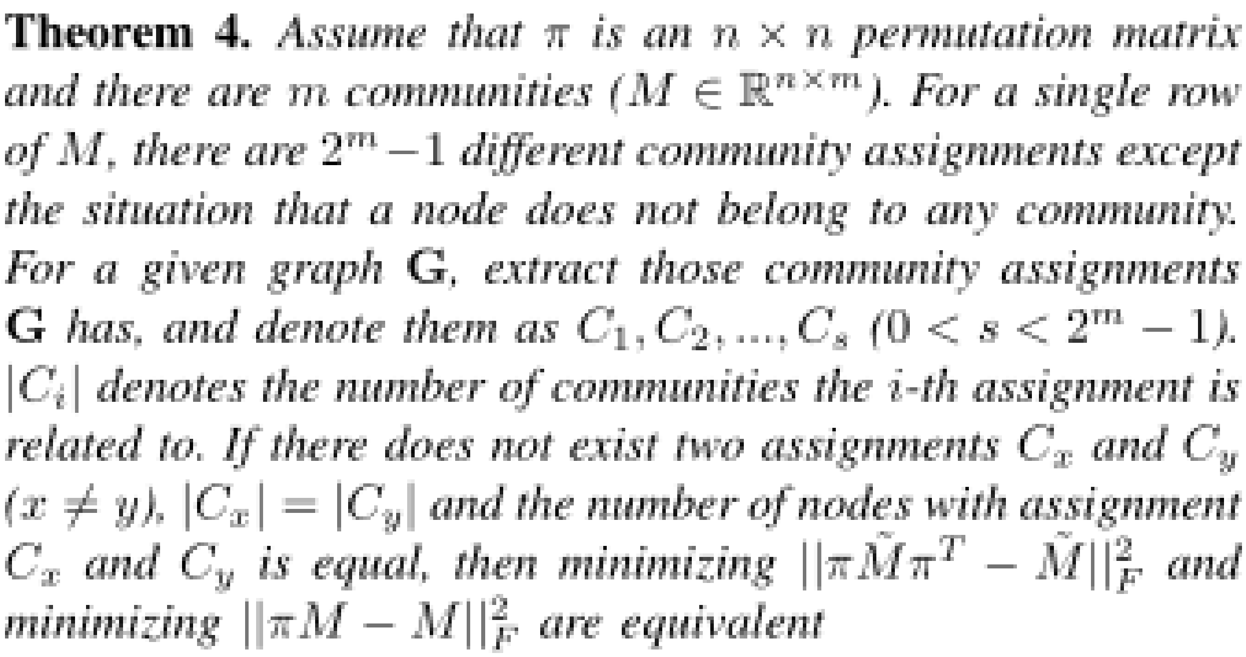


Theorem 3. *Suppose that graph A and B are isomorphic, and the eigenvalue decomposition of A is $A = U\Lambda U^T$, then if $(\mathbf{1}^T U e_i) U^T M M^T U e_j \neq (U^T \mathbf{1} e_i^T) U^T M M^T U e_j$, then there is a unique solution to the edge matching optimization problem.*

4. The equivalent transformation of the community assignment constraint

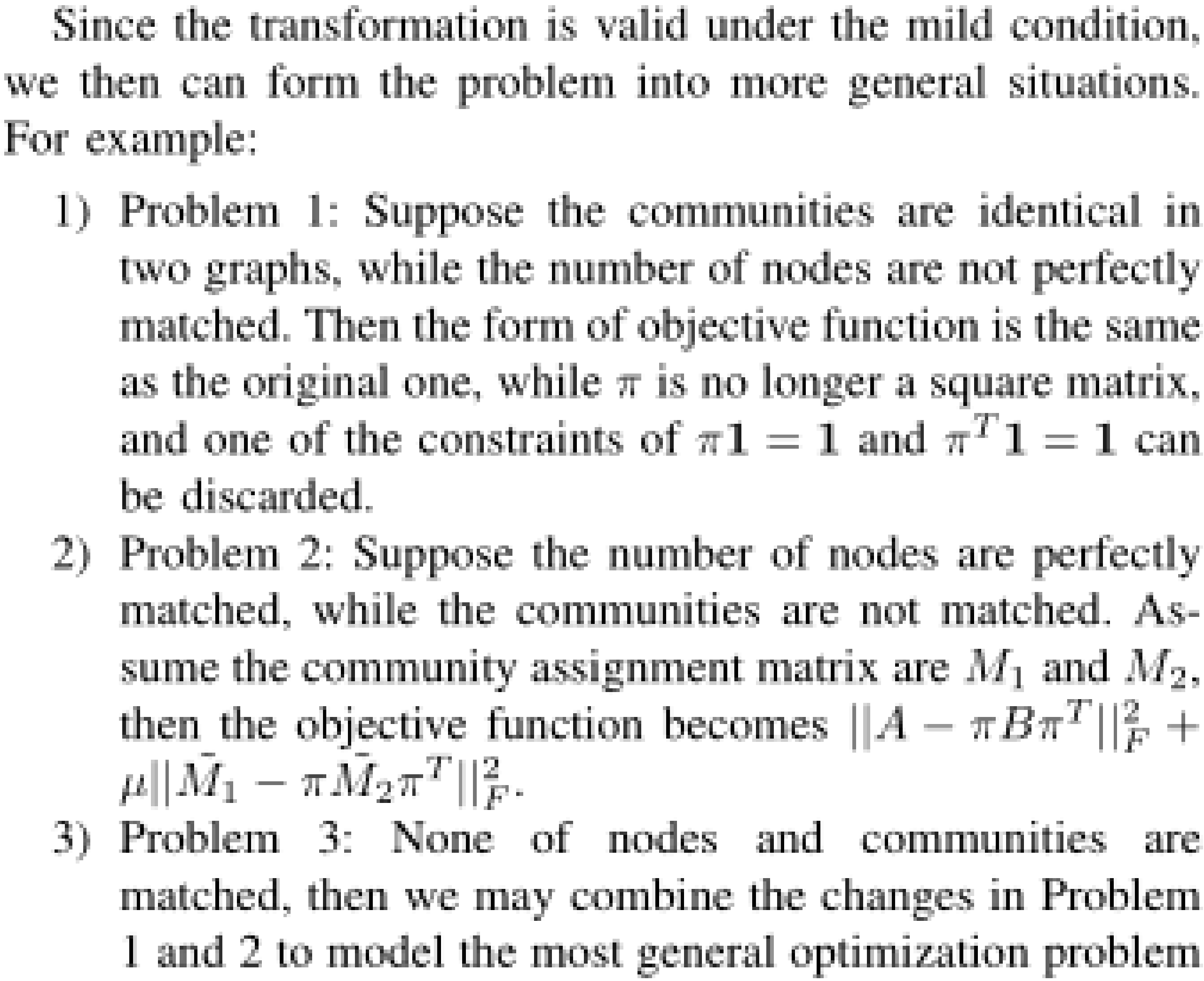
Incentives: $\mu||\pi M - M||_F^2$

The second term of the original objective function is not of a symmetric form, since M is normally not a square matrix. This hinders the extension of the problem into more general problems such as considering different number of nodes or communities, and it makes harder for the theoretical analysis of the algorithm in Section 5 because of its asymmetry.



Theorem 4. *Assume that π is an $n \times n$ permutation matrix and there are m communities ($M \in \mathbb{R}^{n \times m}$). For a single row of M, there are $2^m - 1$ different community assignments except the situation that a node does not belong to any community. For a given graph $\mathbf{G}$, extract those community assignments $\mathbf{G}$ has, and denote them as $C_1, C_2, ..., C_s$ $(0 < s < 2^m - 1)$. $|C_i|$ denotes the number of communities the i-th assignment is related to. If there does not exist two assignments C_x and C_y $(x \neq y)$, $|C_x| = |C_y|$ and the number of nodes with assignment C_x and C_y is equal, then minimizing $||\pi \tilde{M} \pi^T - \tilde{M}||_F^2$ and minimizing $||\pi M - M||_F^2$ are equivalent*

Under the mild condition, we can transform the original second term into the new one, which is of good symmetric property. More importantly, it can be extended to generalize broader problems, which have been discussed separately in previous work.

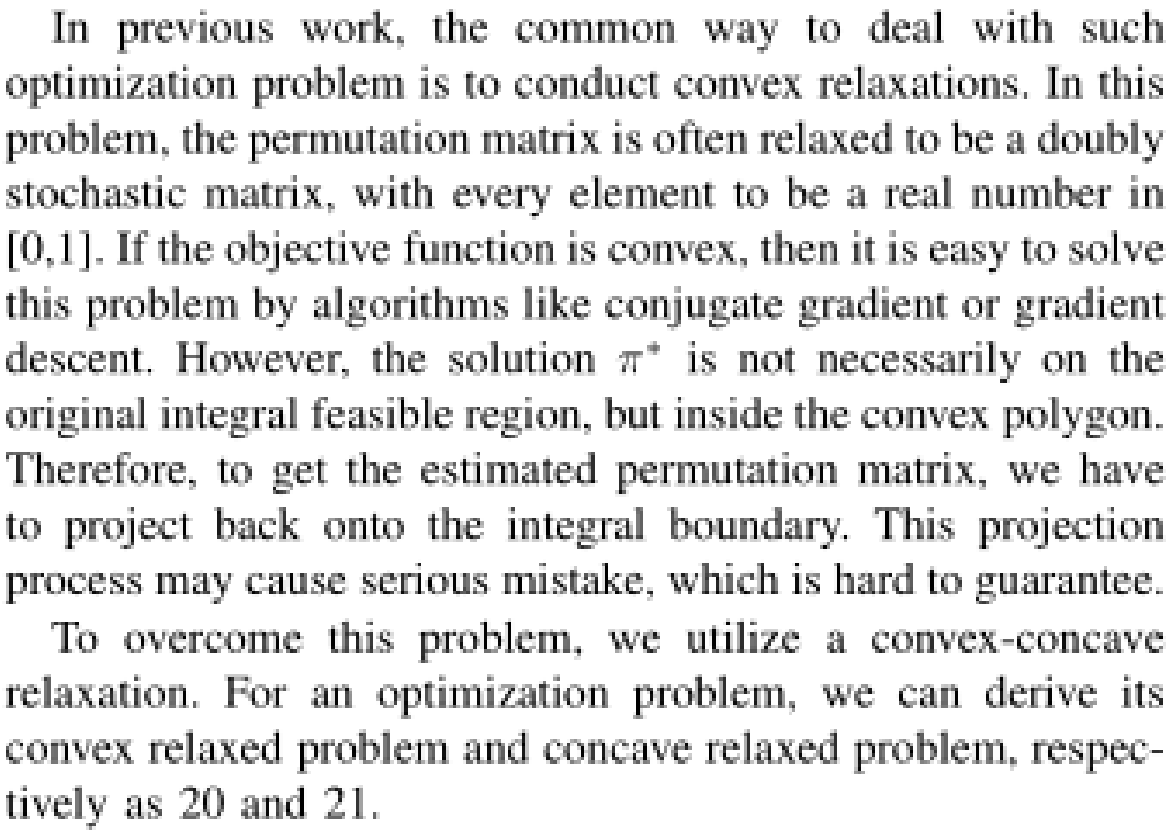


Since the transformation is valid under the mild condition, we then can form the problem into more general situations. For example:

1) Problem 1: Suppose the communities are identical in two graphs, while the number of nodes are not perfectly matched. Then the form of objective function is the same as the original one, while π is no longer a square matrix, and one of the constraints of $\pi\mathbf{1} = \mathbf{1}$ and $\pi^T\mathbf{1} = \mathbf{1}$ can be discarded.
2) Problem 2: Suppose the number of nodes are perfectly matched, while the communities are not matched. Assume the community assignment matrix are M_1 and M_2, then the objective function becomes $||A - \pi B\pi^T||_F^2 + \mu||\bar{M}_1 - \pi\bar{M}_2\pi^T||_F^2$.
3) Problem 3: None of nodes and communities are matched, then we may combine the changes in Problem 1 and 2 to model the most general optimization problem

5. The convex-concave algorithm to solve this problem

Incentives:



In previous work, the common way to deal with such optimization problem is to conduct convex relaxations. In this problem, the permutation matrix is often relaxed to be a doubly stochastic matrix, with every element to be a real number in [0,1]. If the objective function is convex, then it is easy to solve this problem by algorithms like conjugate gradient or gradient descent. However, the solution π^* is not necessarily on the original integral feasible region, but inside the convex polygon. Therefore, to get the estimated permutation matrix, we have to project back onto the integral boundary. This projection process may cause serious mistake, which is hard to guarantee.

To overcome this problem, we utilize a convex-concave relaxation. For an optimization problem, we can derive its convex relaxed problem and concave relaxed problem, respectively as 20 and 21.

$$F_1(\pi) = F_0(\pi), \pi \in \Omega_0, \quad F_1(\pi) \leq F_0(\pi), \pi \in \Omega \tag{20}$$

$$F_2(\pi) = F_0(\pi), \pi \in \Omega_0, \quad F_2(\pi) \geq F_0(\pi), \pi \in \Omega \tag{21}$$

$$F(\pi) = \lambda F_1(\pi) + (1 - \lambda) F_2(\pi) \tag{22}$$

Then we firstly set $\lambda = 0$ and solve the convex problem $F_1(\pi)$, and find its optimal solution π_1. Then we slightly increase λ by $d\lambda$, and solve the new optimization problem $d\lambda F_1(\pi) + (1 - d\lambda) F_2(\pi)$ with the initial value π_1. Then, by gradually increasing λ, the problem becomes much more concave, while the optimal value gradually approaches the boundary, on which the real optimal permutation matrix lies.

A proper way to get the convex relaxation and concave relaxation is as follows.

$$F_1(\pi) = F_0(\pi) - \lambda_{min} \mathbf{tr}(\pi^T \pi - \mathbf{1}\pi) \tag{23}$$

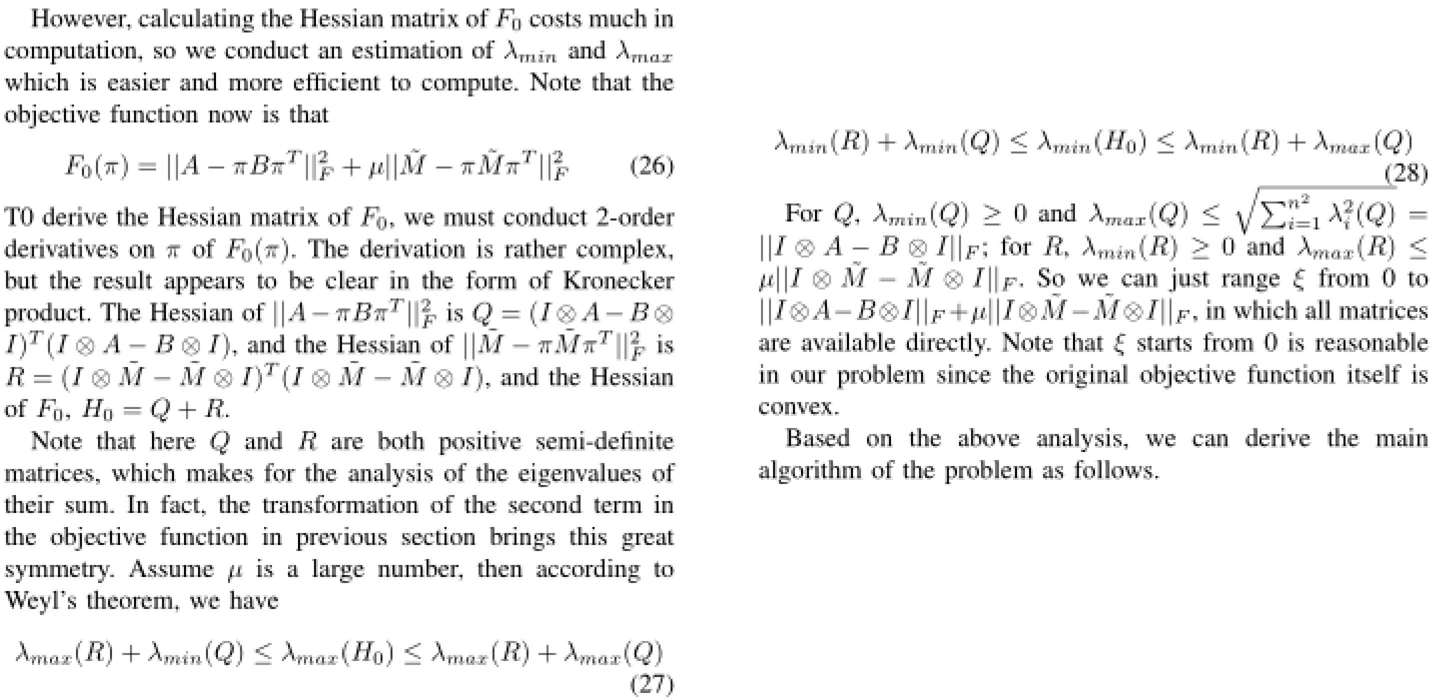
$$F_2(\pi) = F_0(\pi) - \lambda_{max} \mathbf{tr}(\pi^T \pi - \mathbf{1}\pi) \tag{24}$$

where $\mathbf{1}$ is a $n \times n$ matrix with all the elements to be 1, and λ_{min} and λ_{max} are the minimum and maximum eigenvalue of the Hessian matrix of F_0. It is easy to verify they satisfy convex and concave relaxation respectively. Therefore

$$F(\pi) = (1 - \eta) F_1(\pi) + \eta F_2(\pi) = F_0(\pi) - \xi \mathbf{tr}(\pi^T \pi - \mathbf{1}\pi) \tag{25}$$

where $\xi = (1 - \eta)\lambda_{min} + \eta \lambda_{max}$, thus $\xi \in [\lambda_{min}, \lambda_{max}]$.

Estimate the minimum and maximum eigenvalues of the Hessian matrix of the objective function, we can further reduce the computational complexity.



However, calculating the Hessian matrix of F_0 costs much in computation, so we conduct an estimation of λ_{min} and λ_{max} which is easier and more efficient to compute. Note that the objective function now is that

$$F_0(\pi) = ||A - \pi B \pi^T||_F^2 + \mu||\tilde{M} - \pi \tilde{M} \pi^T||_F^2 \qquad (26)$$

T0 derive the Hessian matrix of F_0, we must conduct 2-order derivatives on π of $F_0(\pi)$. The derivation is rather complex, but the result appears to be clear in the form of Kronecker product. The Hessian of $||A - \pi B \pi^T||_F^2$ is $Q = (I \otimes A - B \otimes I)^T (I \otimes A - B \otimes I)$, and the Hessian of $||\tilde{M} - \pi \tilde{M} \pi^T||_F^2$ is $R = (I \otimes \tilde{M} - \tilde{M} \otimes I)^T (I \otimes \tilde{M} - \tilde{M} \otimes I)$, and the Hessian of F_0, $H_0 = Q + R$.

Note that here Q and R are both positive semi-definite matrices, which makes for the analysis of the eigenvalues of their sum. In fact, the transformation of the second term in the objective function in previous section brings this great symmetry. Assume μ is a large number, then according to Weyl's theorem, we have

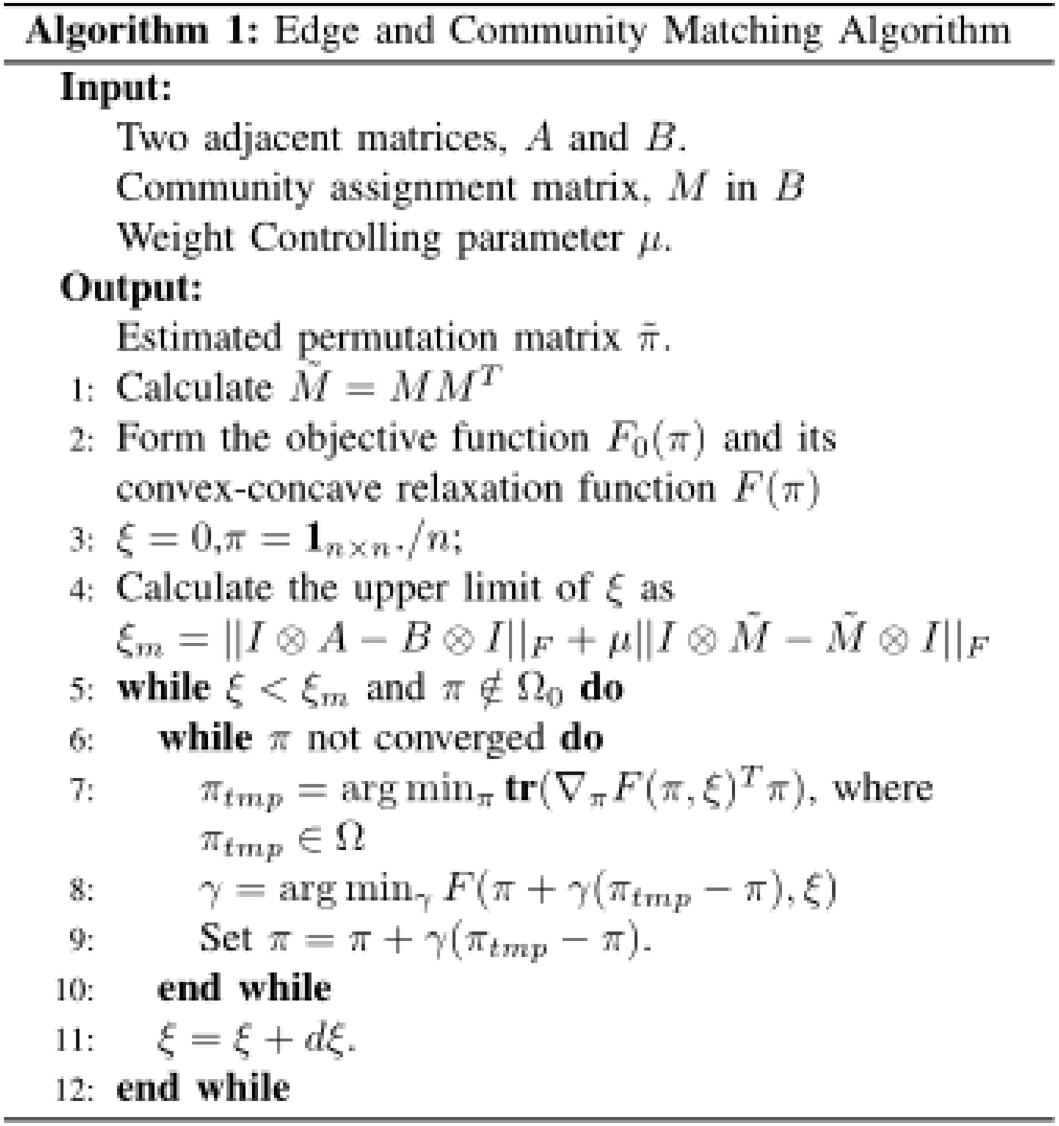
$$\lambda_{max}(R) + \lambda_{min}(Q) \leq \lambda_{max}(H_0) \leq \lambda_{max}(R) + \lambda_{max}(Q) \qquad (27)$$

$$\lambda_{min}(R) + \lambda_{min}(Q) \leq \lambda_{min}(H_0) \leq \lambda_{min}(R) + \lambda_{max}(Q) \qquad (28)$$

For Q, $\lambda_{min}(Q) \geq 0$ and $\lambda_{max}(Q) \leq \sqrt{\sum_{i=1}^{n^2} \lambda_i^2(Q)} = ||I \otimes A - B \otimes I||_F$; for R, $\lambda_{min}(R) \geq 0$ and $\lambda_{max}(R) \leq \mu||I \otimes \tilde{M} - \tilde{M} \otimes I||_F$. So we can just range ξ from 0 to $||I \otimes A - B \otimes I||_F + \mu||I \otimes \tilde{M} - \tilde{M} \otimes I||_F$, in which all matrices are available directly. Note that ξ starts from 0 is reasonable in our problem since the original objective function itself is convex.

Based on the above analysis, we can derive the main algorithm of the problem as follows.

After transformation, we get our algorithm.



Algorithm 1: Edge and Community Matching Algorithm

```
Input:
    Two adjacent matrices, $A$ and $B$.
    Community assignment matrix, $M$ in $B$
    Weight Controlling parameter $\mu$.
Output:
    Estimated permutation matrix $\bar{\pi}$.
Calculate $\tilde{M} = MM^T$
Form the objective function $F_0(\pi)$ and its
convex-concave relaxation function $F(\pi)$
$\xi = 0, \pi = \mathbf{1}_{n\times n}./n$;
Calculate the upper limit of $\xi$ as
$\xi_m = ||I \otimes A - B \otimes I||_F + \mu||I \otimes \tilde{M} - \tilde{M} \otimes I||_F$
while $\xi < \xi_m$ and $\pi \notin \Omega_0$ do
    while $\pi$ not converged do
        $\pi_{tmp} = \arg\min_\pi \mathbf{tr}(\nabla_\pi F(\pi,\xi)^T \pi)$, where
        $\pi_{tmp} \in \Omega$
        $\gamma = \arg\min_\gamma F(\pi + \gamma(\pi_{tmp} - \pi), \xi)$
        Set $\pi = \pi + \gamma(\pi_{tmp} - \pi)$.
    end while
    $\xi = \xi + d\xi$.
end while
```

Future work:

1. Analyze how the overlapping communities will affect the accuracy, and modify the objective function or algorithm to harness the overlapping property.

2. Analyze the performance of the convex-concave algorithm.

3. Conduct experiments based on real data

Thanks!