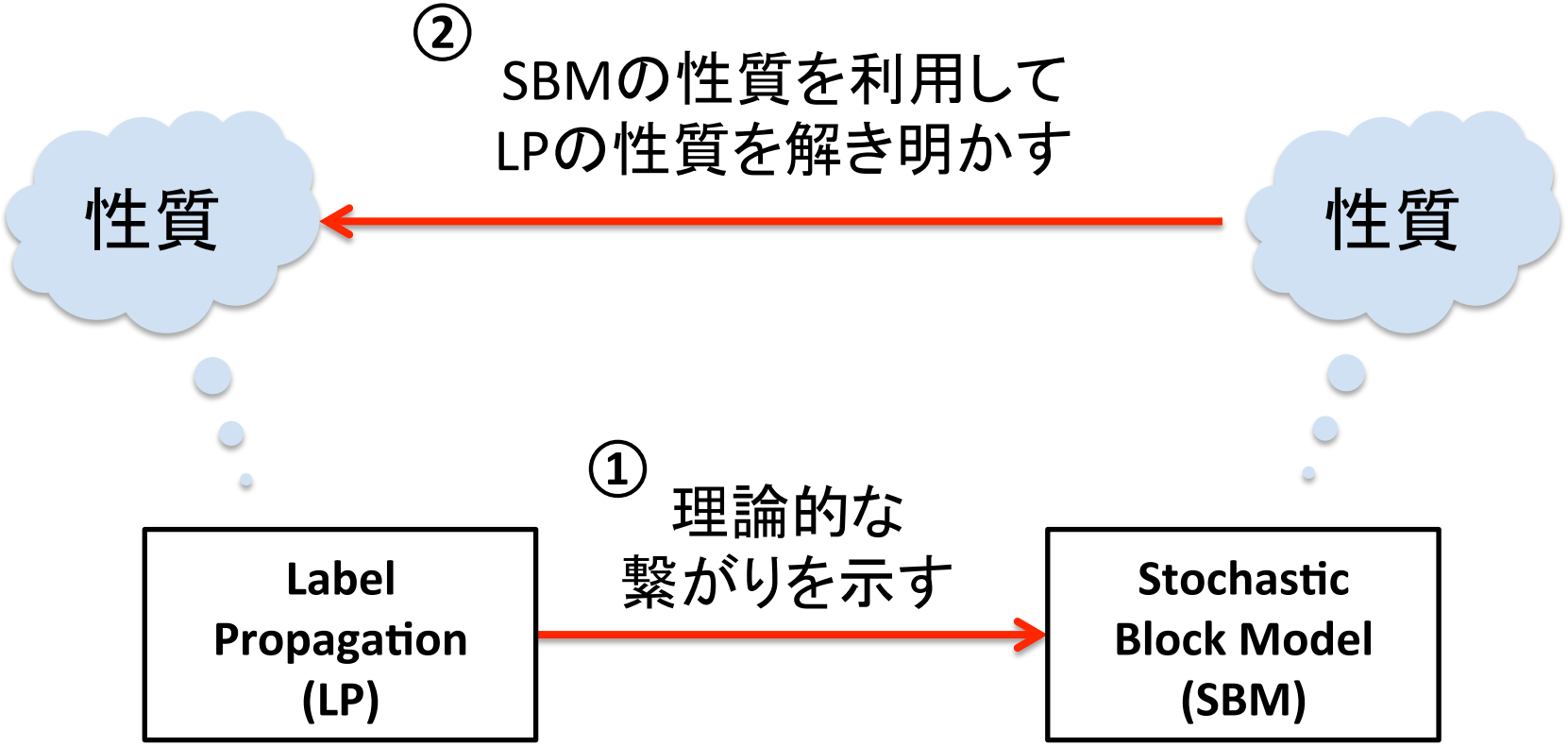
When Does Label Propagation Fail? A View from a Network Generative Model

[IJCAI'17]

Yuto Yamaguchi and Kohei Hayashi

AIST

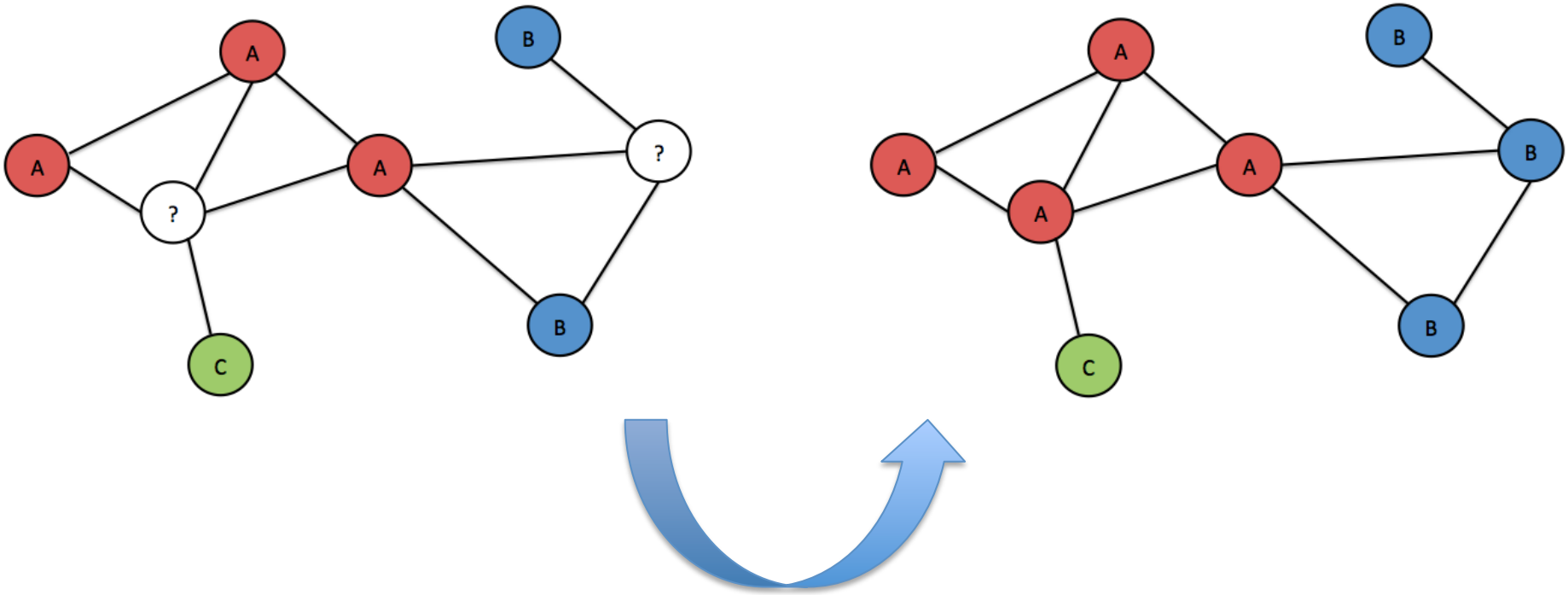
ネットワーク生成モデルの見地からLPの性質を解析する



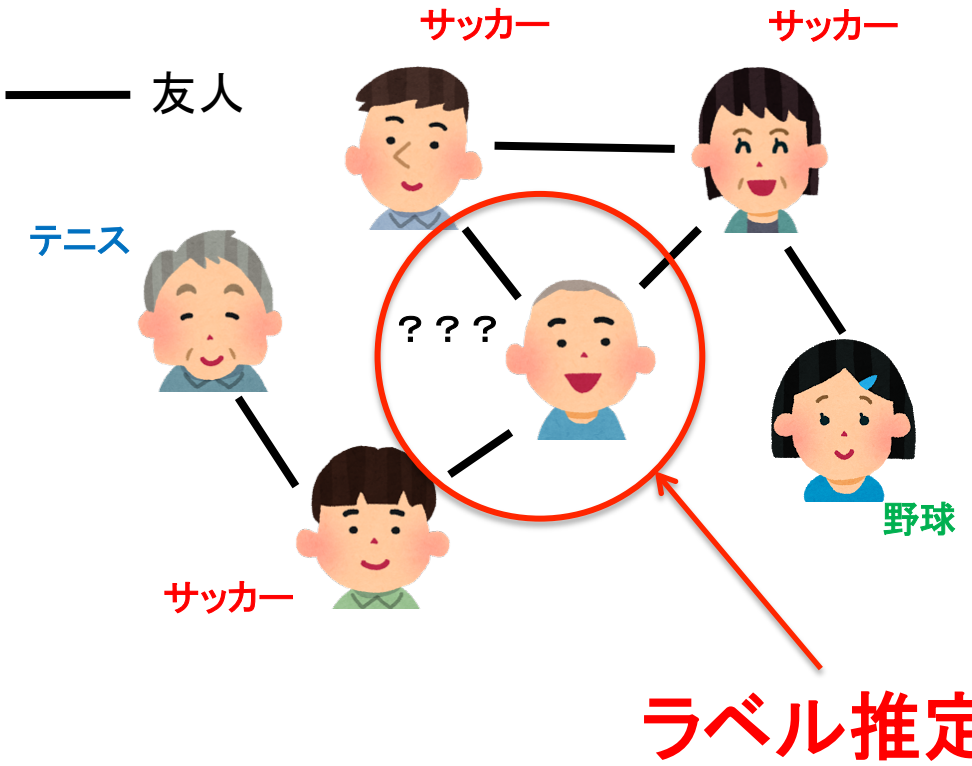
問題定義｜ラベル分類

Given 部分的にラベルが付いた無向グラフ

Find 全部のラベル



例｜SNSユーザの属性推定

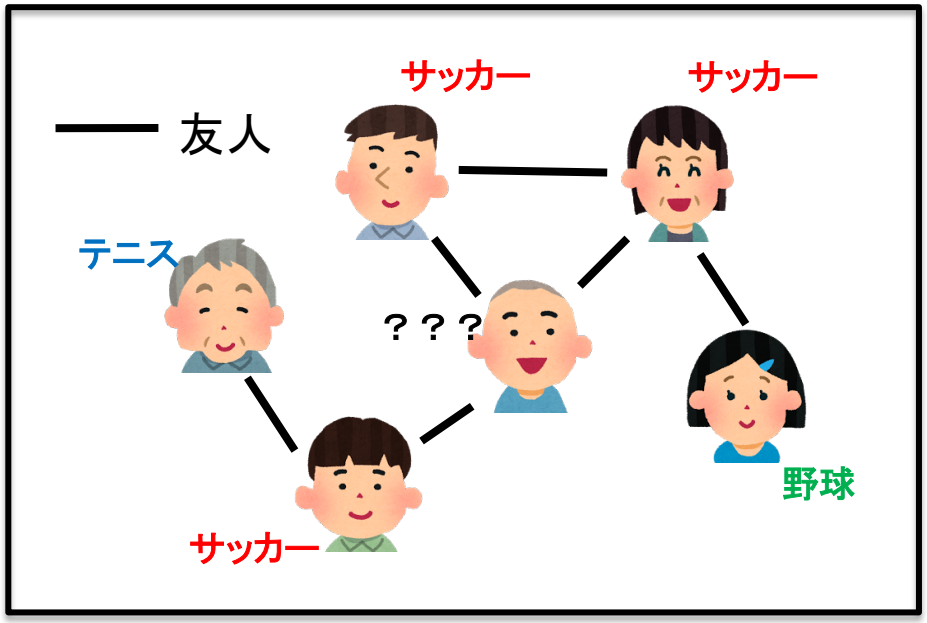


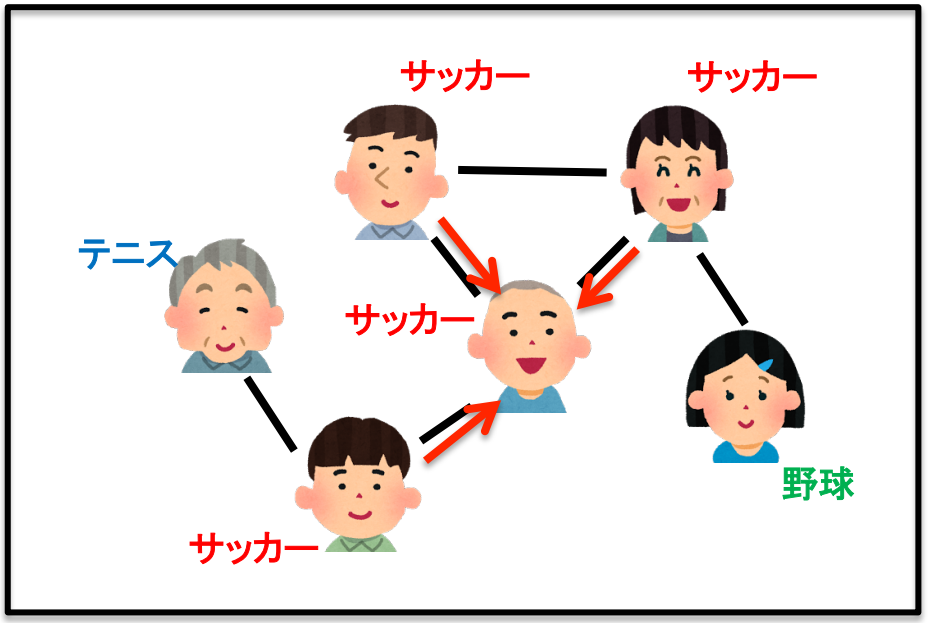
- 真ん中の人の趣味は何？

Label Propagation (1/2)

[Zhu+, 03], [Zhou+, 03], など

既知のラベルを伝播させる





Label Propagation (2/2)

[Zhu+, 03], [Zhou+, 03], など

隣接行列 X と既知のラベル Y が与えられたとき
Q を最大にする F = { fi } を求める

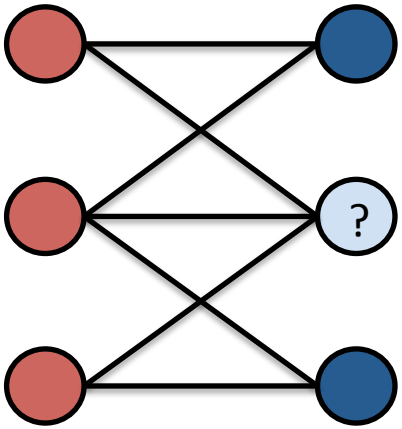
$$Q(F;X,Y,\lambda)=\frac{1}{2}\sum_{i=1}^{N}\|f_i-y_i\|_2^2+\frac{\lambda}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}x_{ij}\|f_i-f_j\|_2^2$$

既知のラベルと推定ラベルは
出来るだけ等しい

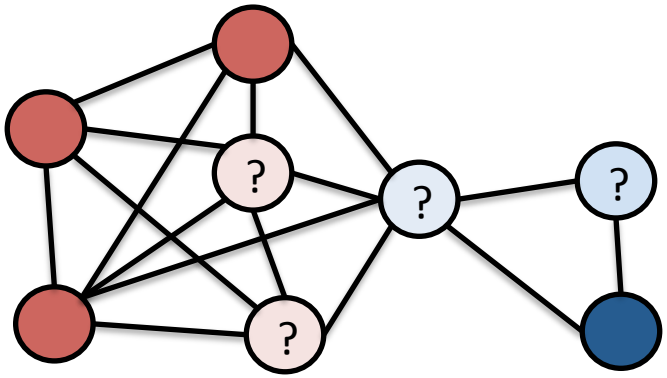
隣り合うノードのラベルは
出来るだけ等しい

LP がうまくいかないケース

同一のラベルが隣り合わないケース



ラベルの割合が偏っているケース

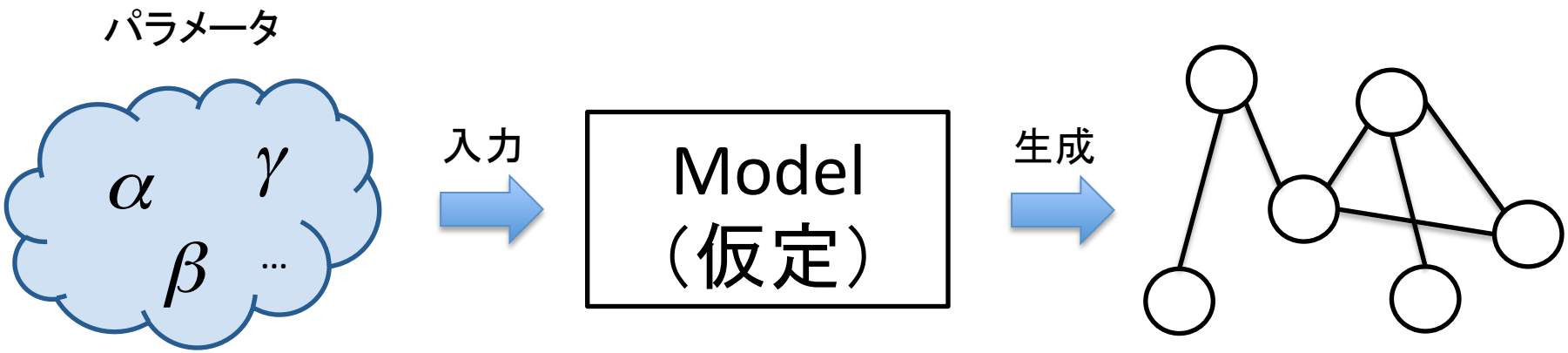


Q. なんでうまくいかない？
Q. 他にはどういうケースがある？

本研究の成果

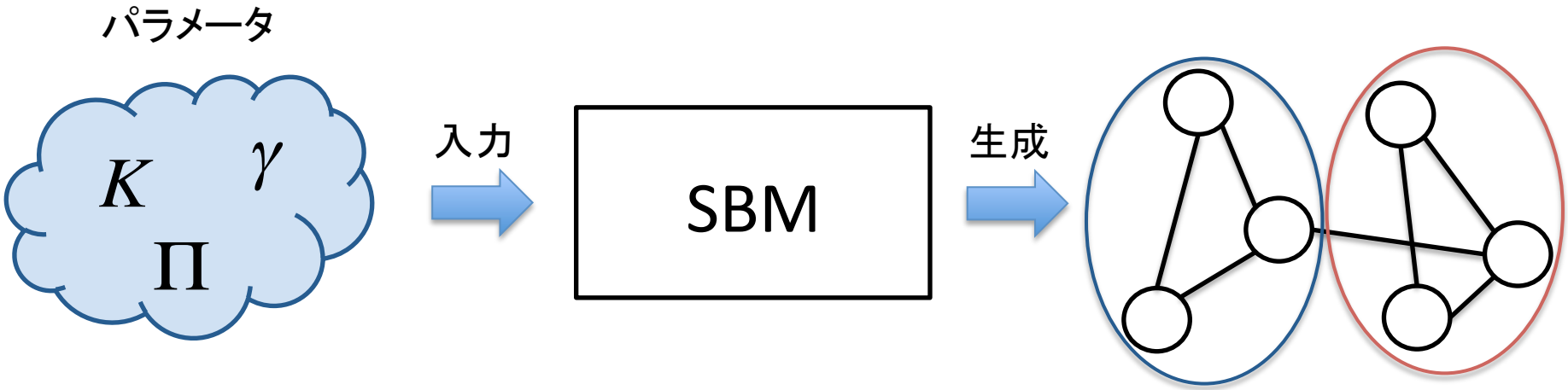
- LP と SBM の理論的なつながりを示す
- ネットワーク生成モデルの見地からLPの性質を解析
- LPが（暗黙的に）置いている仮定を示す

Network Generative Model



入力されたパラメータとモデルの仮定をもとに
実際のネットワーク（データ）を生成する

Stochastic Block Model (1/2)



SBM の仮定

- 全てのノードは K 個あるクラスタのうち**必ずどれか一つ**に属す
  - 属す確率は K 次元確率ベクトル γ によって定まる
- あるノード i と j の間にエッジが存在する確率は i と j が**属すクラスタのみ**によって定まる
  - 接続確率は K x K 行列 Π によって定まる

Stochastic Block Model (2/2)

生成過程

多項分布

パラメータ

パラメータ

ベルヌーイ分布

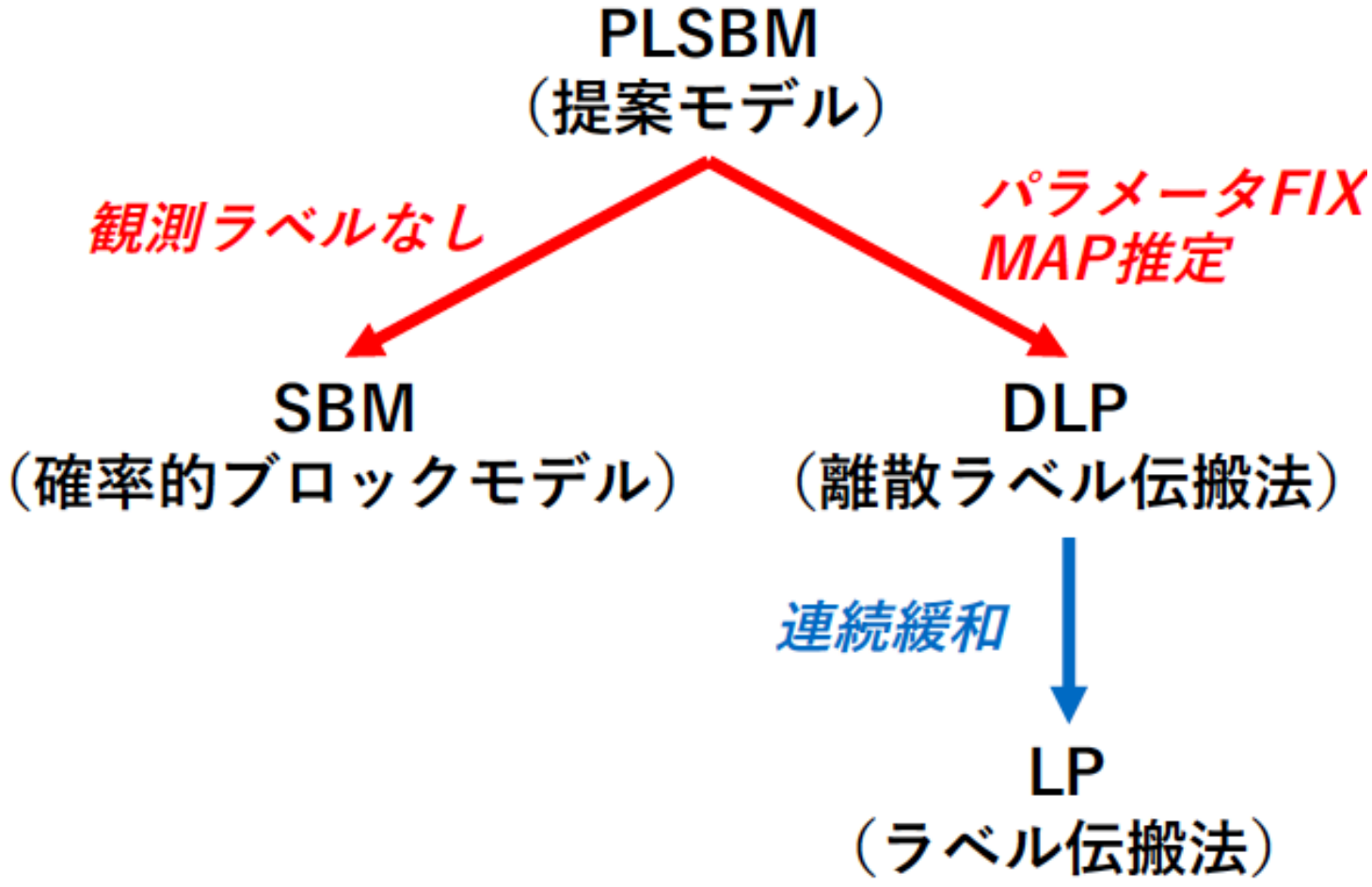
① • For each node $i = 1, \ldots, N$
  – Generate $z_i \sim Mult(\cdot|\gamma)$

② • For each node pair (i, j)
  – Generate $x_{ij} \sim Bern(\cdot|z_i^{\mathrm{T}} \mathbf{\Pi} z_j)$

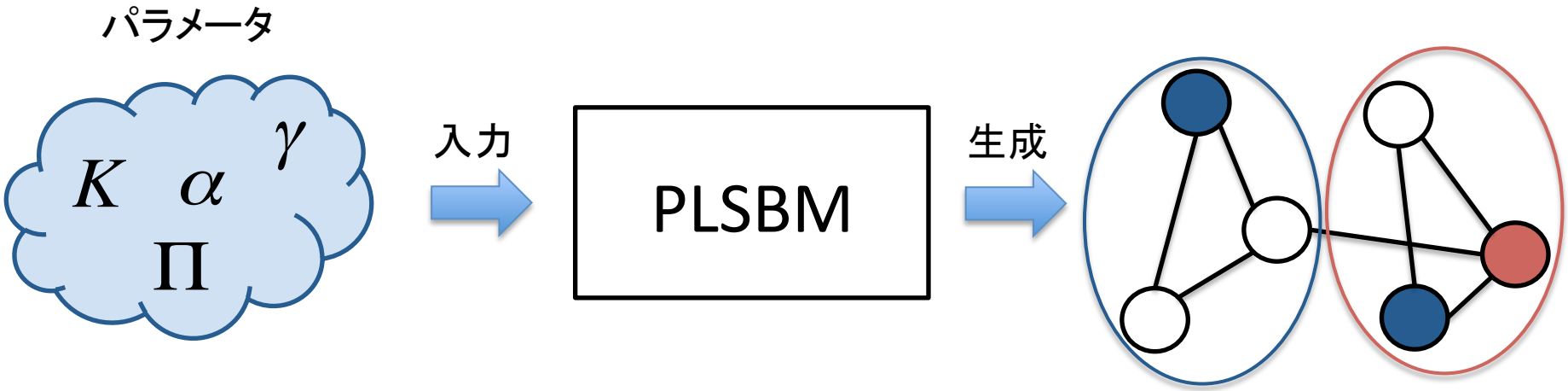
①: 各ノードのクラスタ割り当て z を生成

②: 隣接行列 x の要素を生成

LP と SBM の理論的関係



Partially Labeled SBM (PLSBM)



PLSBM の仮定

- SBM の仮定
- 一部のノードはクラスタ割り当てとは別に “ラベル” を持つ
  - ラベルとクラスタ割り当ては α が大きいほど一致する

Partially Labeled SBM (PLSBM)

生成過程

①
- For each node $i = 1, \ldots, N$
  - Generate $z_i \sim Mult(\cdot|\gamma)$

②
- For each labeled node $i = 1, \ldots, N_L$
  - Generate $y_i \sim Mult(\cdot|\underline{B}z_i)$

③
- For each node pair (i, j)
  - Generate $x_{ij} \sim Bern(\cdot|z_i^{\mathrm{T}}\Pi z_j)$

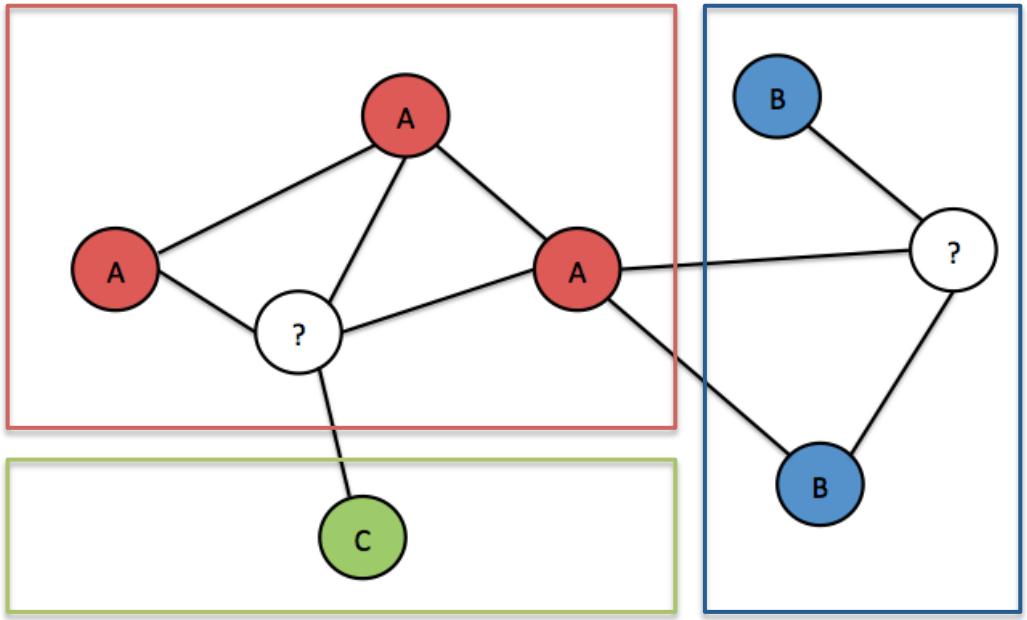
SBMに
②を追加

パラメータ α
によって計算される

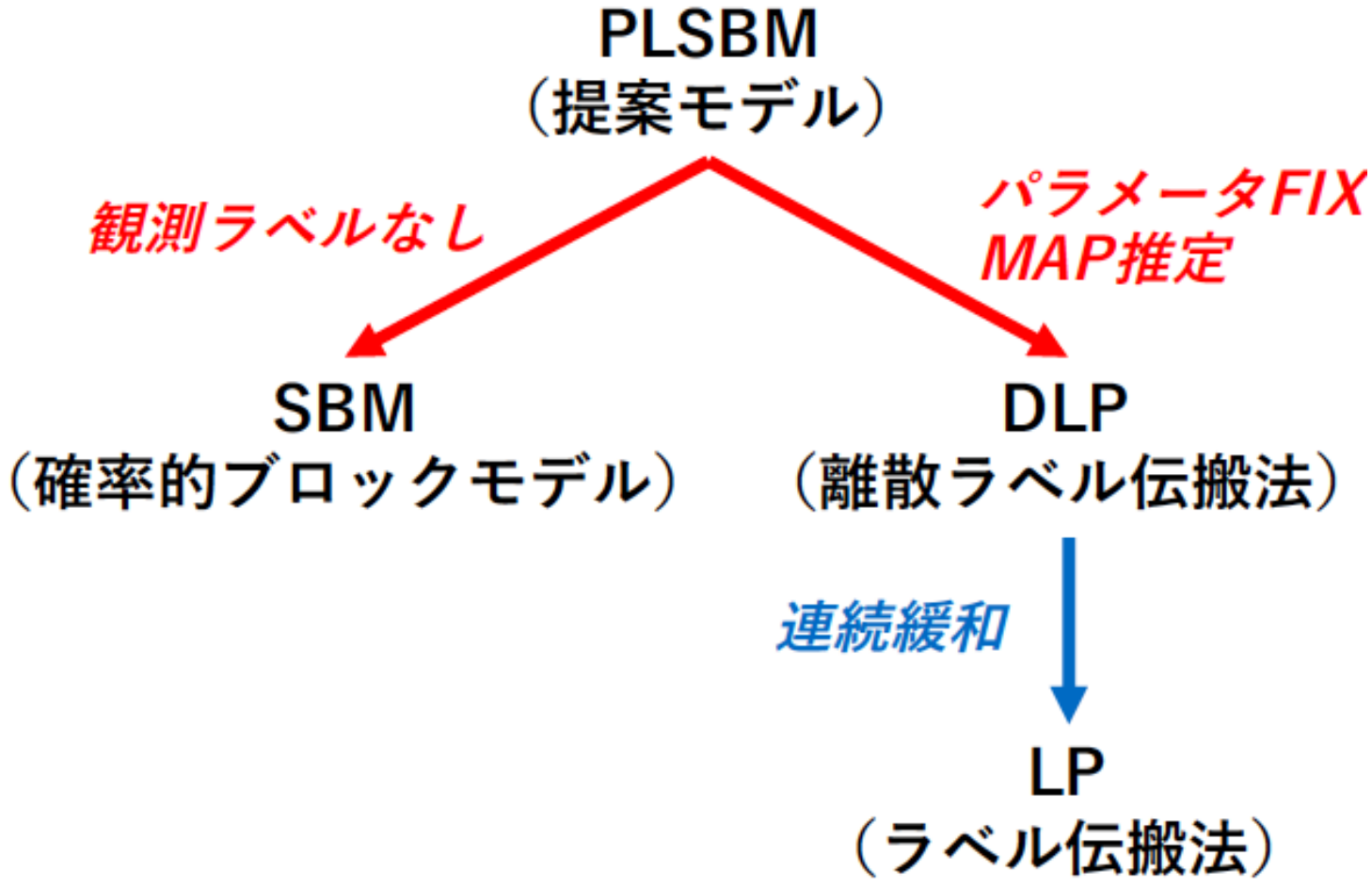
②：各ノードの “ラベル” を生成
（クラスタ割り当てをもとに生成される）

PLSBM でラベル推定

隣接行列 x と(一部の)ラベル y が与えられた上で z を推定



LP と SBM の理論的関係



Discrete Label Propagation

隣接行列 X と(一部の)ラベル y が与えられた上で z を推定

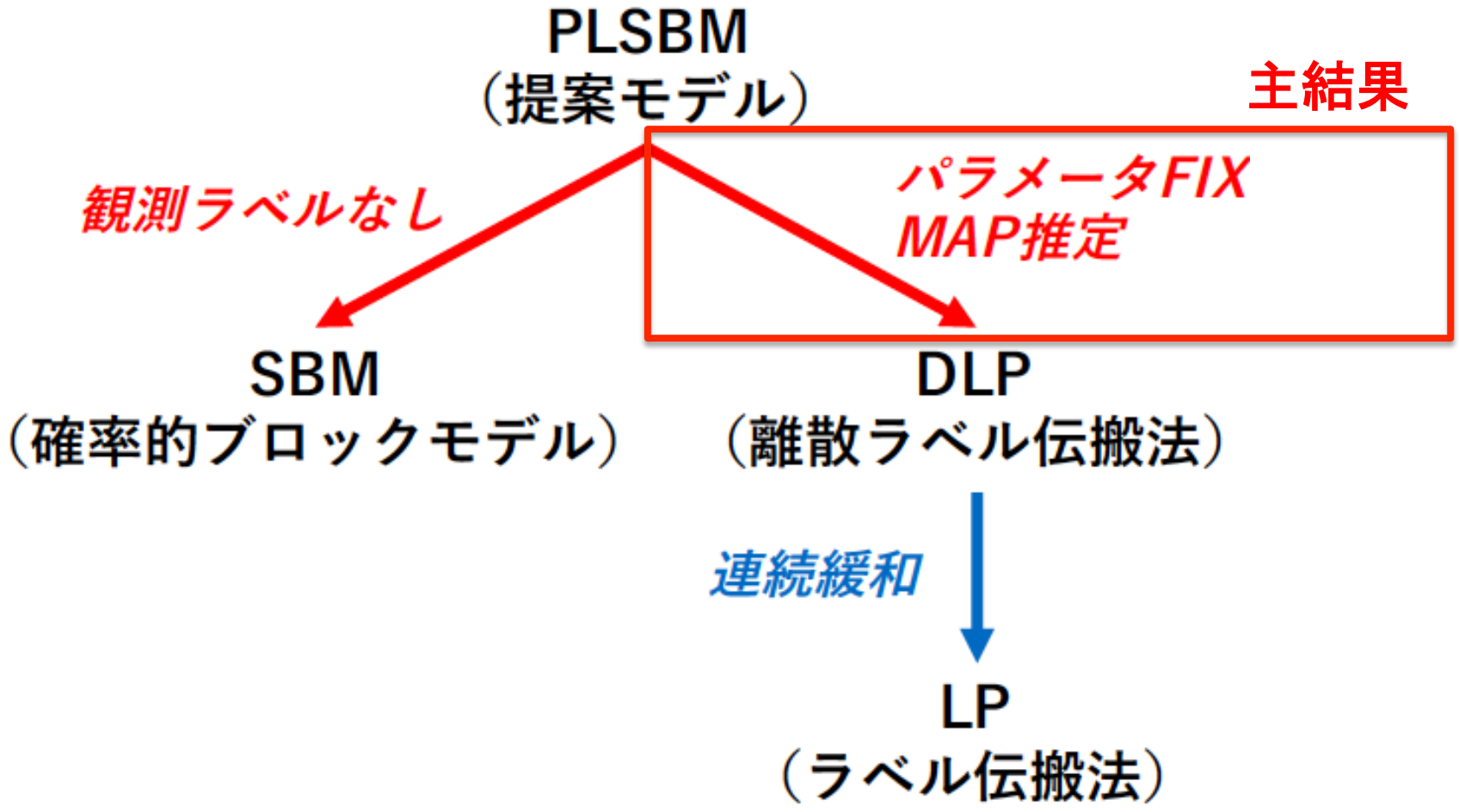
DLPの目的関数（LPと同じ）

$$Q(Z;X,Y,\lambda)=\frac{1}{2}\sum_{i=1}^{N}\|z_i-y_i\|_2^2+\frac{\lambda}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}x_{ij}\|z_i-z_j\|_2^2$$

Q を最大化する Z を求める（組み合わせ最適化）

➔ **Label propagation は Z を連続緩和して解いている**

LP と SBM の理論的関係



主結果: PLSBM と DLP の関係

以下の条件が満たされるとき、
DLP の解と PLSBM における z の MAP 推定は一致

- Condition 1: $\gamma_k = 1/K \; for \; all \; k,$
- Condition 2: $\mathbf{\Pi} = \mu \boldsymbol{I} + \nu(11^{\mathrm{T}} - \boldsymbol{I}),$
- Condition 3: $\lambda \ln \frac{\alpha(K-1)}{1-\alpha} = \ln \frac{\mu(1-\nu)}{\nu(1-\mu)}$
- Condition 4:（略）

Condition 1

$$\gamma_k = 1/K \; for \; all \; k,$$

言っていること

- γ の値は全ての k において等しい

Implication（DLPの暗黙的仮定）

- ネットワーク上のラベルの割合は一定

Condition 2

$$\mathbf{\Pi} = \mu \boldsymbol{I} + \nu(\mathbf{1}\mathbf{1}^{\mathrm{T}} - \boldsymbol{I})$$

言っていること

- Π の対角成分の値はそれぞれ等しい
- Π の非対角成分の値はそれぞれ等しい

Implication（DLPの暗黙的仮定）

- 各クラスタ内のエッジの密度はそれぞれ等しい
- 各クラスタ間のエッジの密度はそれぞれ等しい

Condition 3

$$\lambda \ln \frac{\alpha(K-1)}{1-\alpha} = \ln \frac{\mu(1-\nu)}{\nu(1-\mu)}$$

言っていること

- $\lambda > 0$ かつ $\alpha > 1/K$ ⇒ $\mu > \nu$

Implication（DLPの暗黙的仮定）

- （LPは必ず λ が正なので）
  既知のラベルがランダム以上に合っていれば
  クラスタ内のエッジ密度はクラスタ間よりも大きい

まとめ

- Label Propagation と Stochastic Block Model の理論的なつながりを示した
  - Label Propagation をネットワーク生成モデルとして解釈
- Label Propagation が暗黙的に置いている仮定を示した