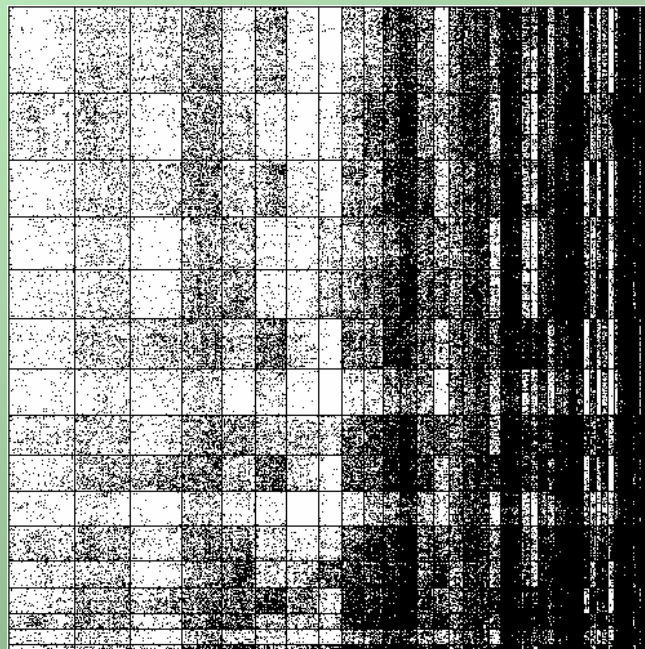


複雑ネットワークと機械学習 無限関係モデルとその周辺

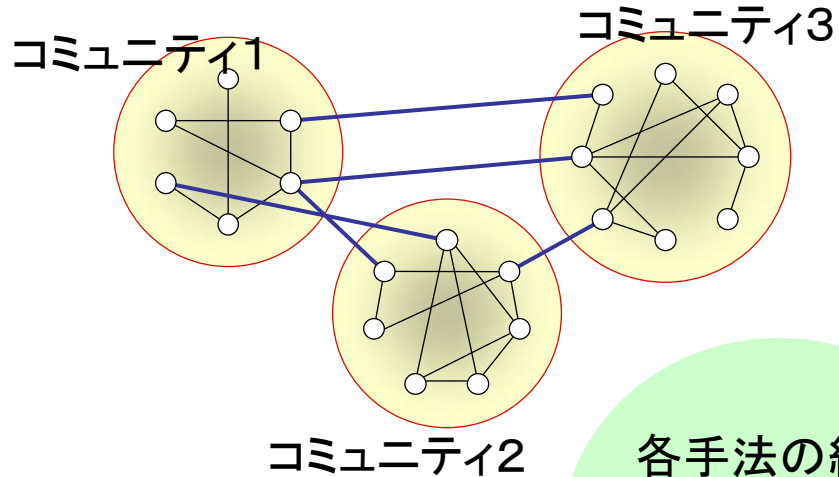


NTTコミュニケーション科学基礎研究所
創発環境研究グループ

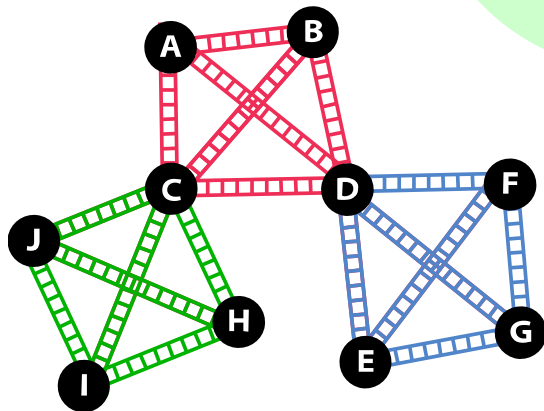
山田 武士

関係データやネットワークのクラスタリング

ネットワークからのコミュニティ抽出

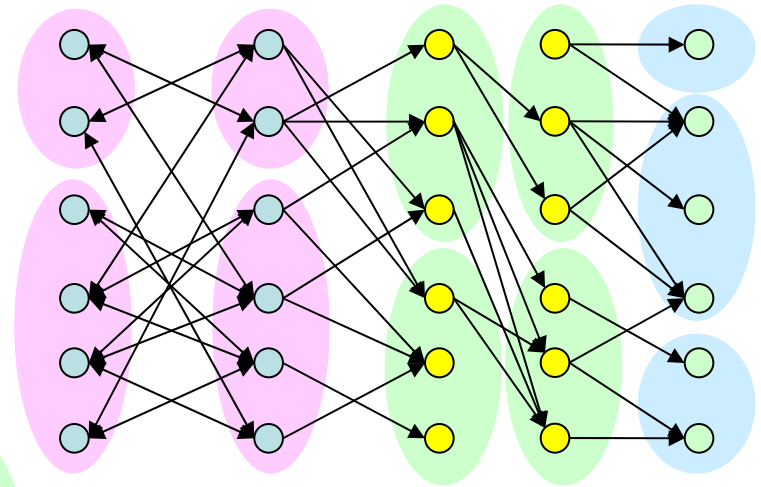


リンクのクラスタリング

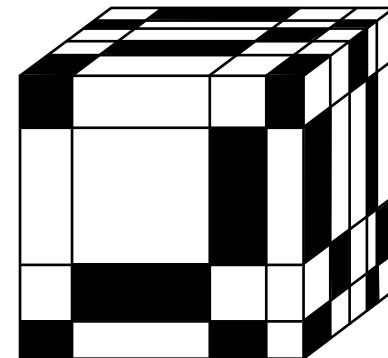


各手法の紹介
と
手法間の
関係の整理

より複雑な関係データのクラスタリング



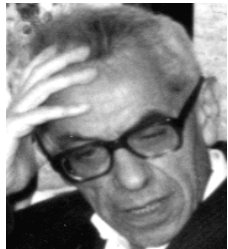
多重クラスタリング
2ドメインの場合は Co-Clustering
(stochastic block models)



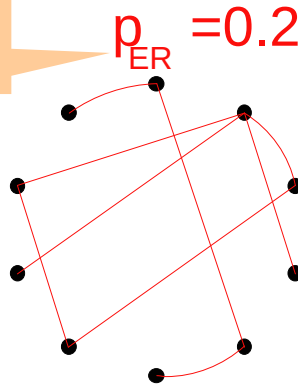
スケールフリーネットワーク

k : 次数 (ノードのリンク数)、 $P(k)$: 次数 k のノード数の割合 (確率)

一定の確率で
リンクを張る

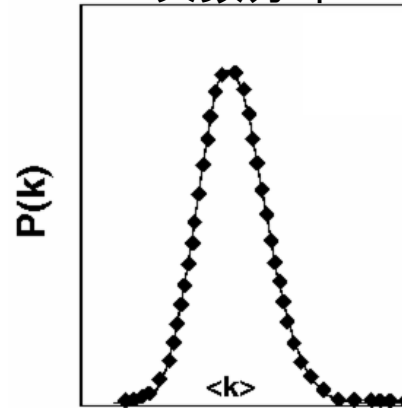


Erdős

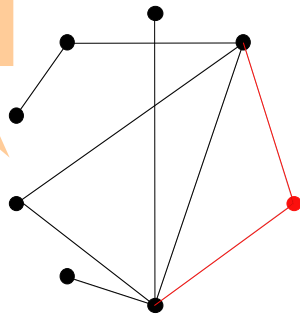


Erdős-Rényi ランダムネットワーク

次数分布

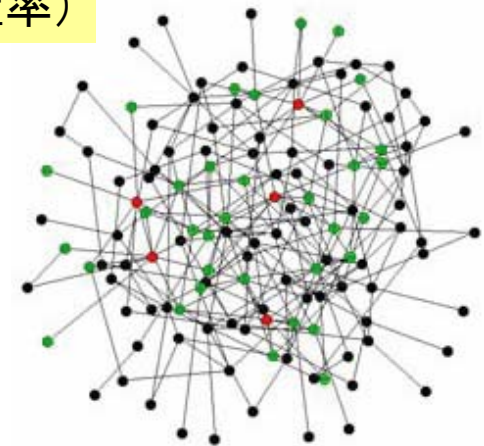
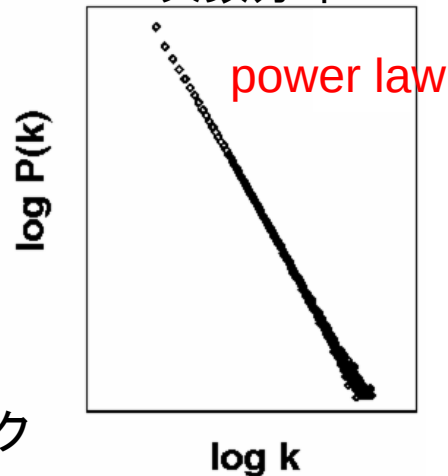


次数に比例する
確率でリンクを張る

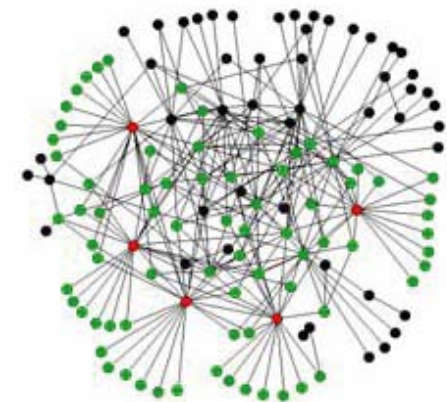


スケールフリーネットワーク

次数分布



もっとも次数の高い5ノード(赤)と繋がっている(緑)のは27%



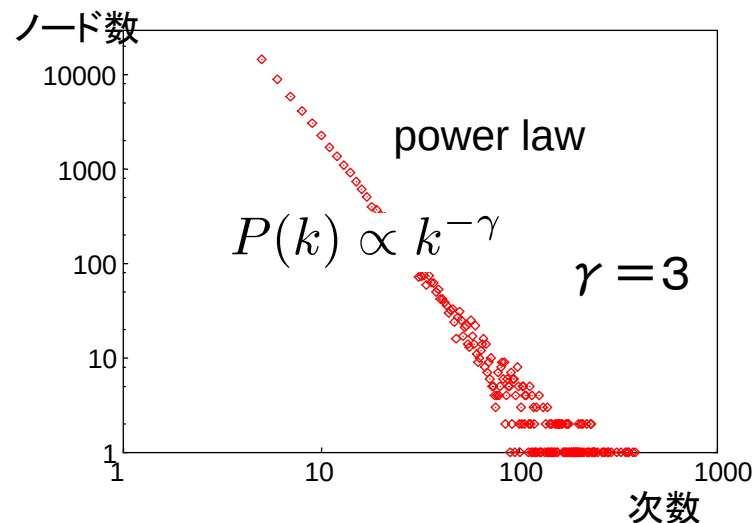
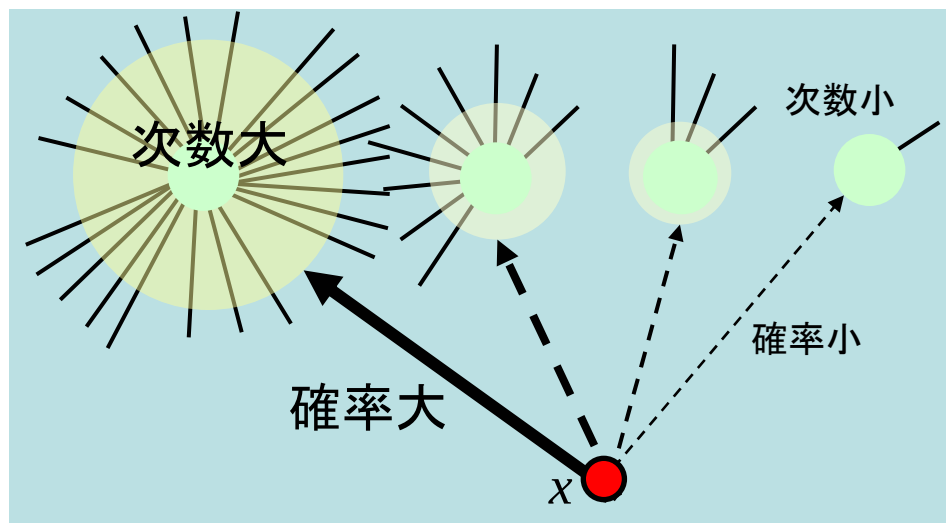
もっとも次数の高い5ノード(赤)と繋がっている(緑)のは60%

優先選択モデル(Preferential Attachment)

[Barabasi & Albert 1999]

- 現在のネットワークに新たなノード x を追加
- 次数(リンク数)に比例する確率で、既存のノードをランダムに選択し、 x からリンクを張る、これを m 回行う(x からのリンク m 本)

⇒ ほぼ係数 $\gamma = 3$ のべき則(power law)に従うネットワークを生成

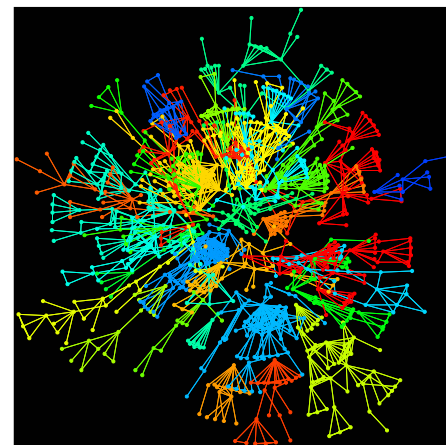
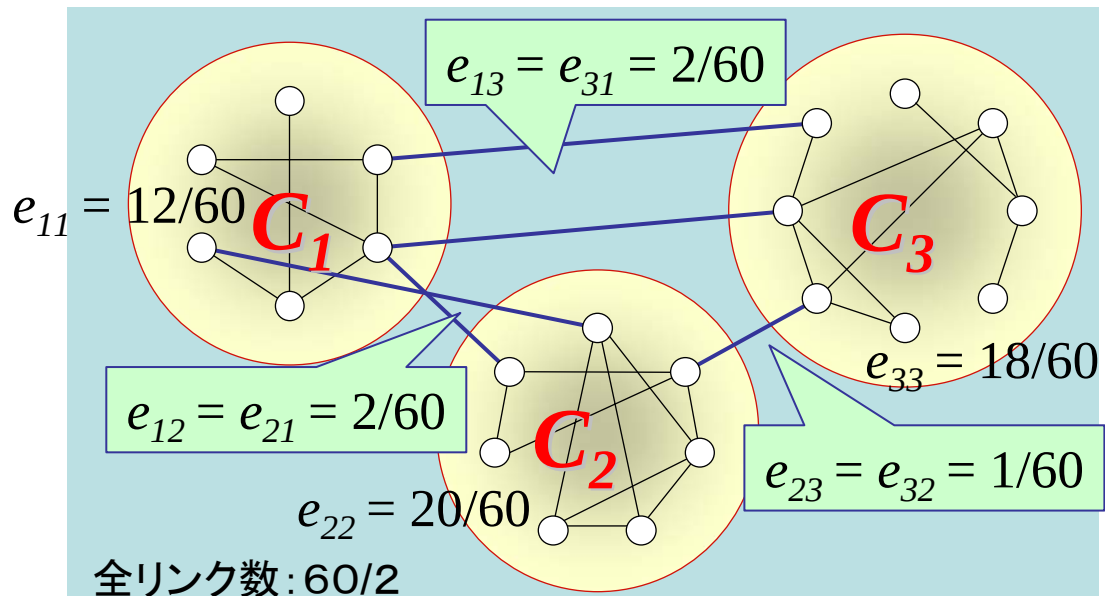


- すでにたくさんリンクを持つノードほど、新たなリンクを張りやすい
- リンクが密な部分、疎な部分が混在⇒「構造」を持つネットワーク

「構造」抽出

Newman の Network Modularity

同一コミュニティ内のリンクは密、コミュニティ間のリンクは疎



C_i, C_j 間のリンク数の割合

相手コミュニティを区別せずリンクを張る場合の C_i, C_j 間のリンク数の割合の期待値

$$a_i a_j \Leftrightarrow e_{ij}$$

コミュニティがリンクで特徴付けられない
⇒両者はほぼ一致

$$\begin{cases} e_{ij} = \#links(i \leftrightarrow j) / \#links(all) \\ a_i = \sum_j e_{ij} \end{cases}$$

Modularity (モジュール度)

$$Q = \sum_i (e_{ii} - a_i^2) \rightarrow \text{最大}$$

コミュニティ内リンクに着目

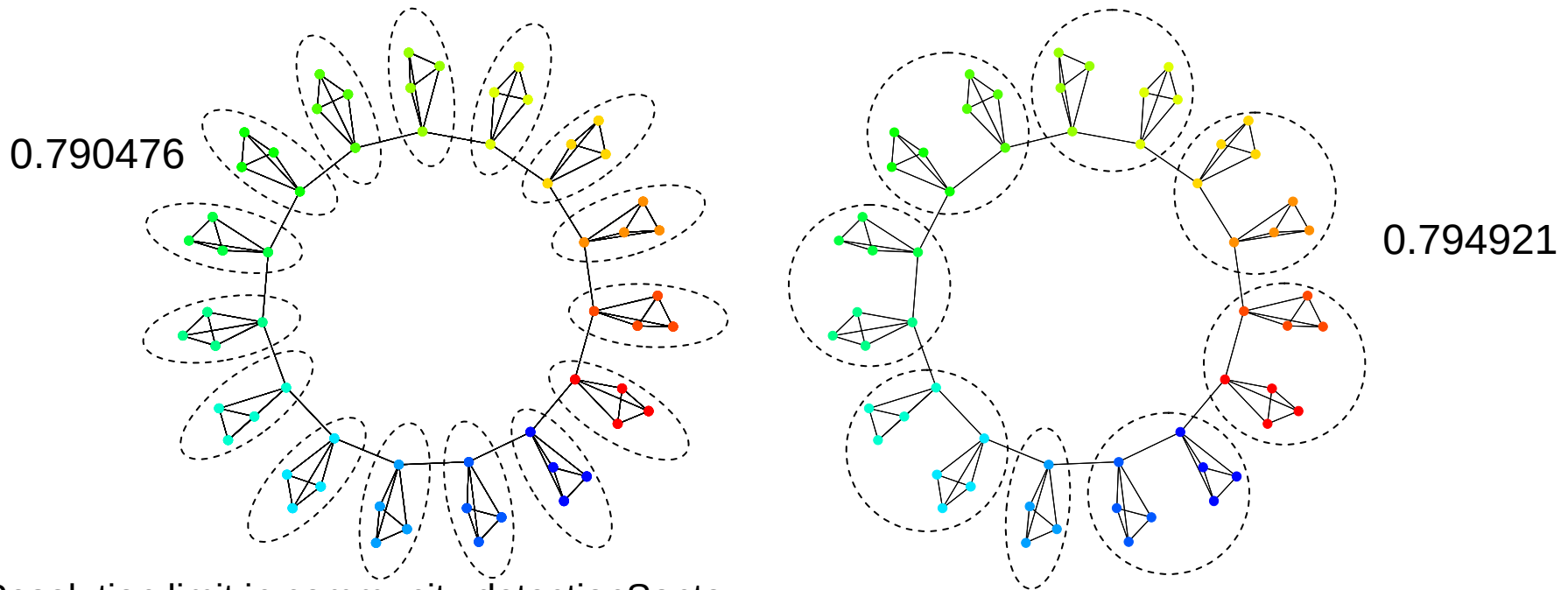
Network Modularity の特徴と問題点

特徴

- 同一クラスタ(コミュニティ)内は、クラスタ間に比べて、相対的に、より密に結合している。
- クラスタ外へのリンクの張り方は似ていなくても良い

問題点

- 確率モデルではない(ノイズ等を考慮していない)
- resolution limit の問題がある



Resolution limit in community detection Santo,
Fortunato and Marc Barthélemy, Proc. Natl. Acad. Sci. USA 104, 36–41 (2007).

Mixture Model によるノードクラスタリング

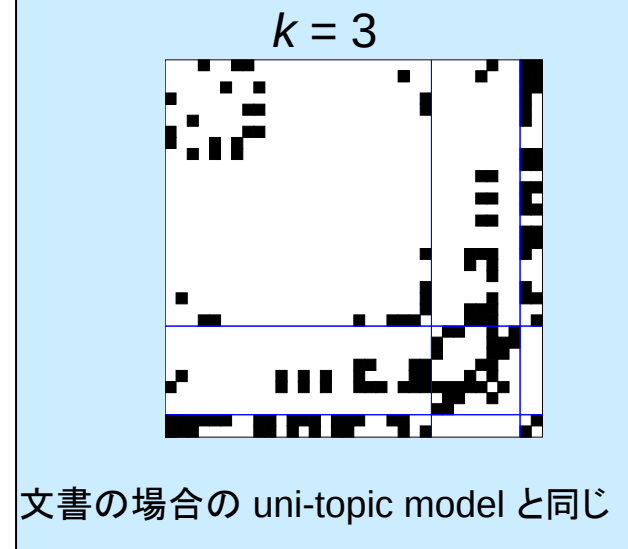
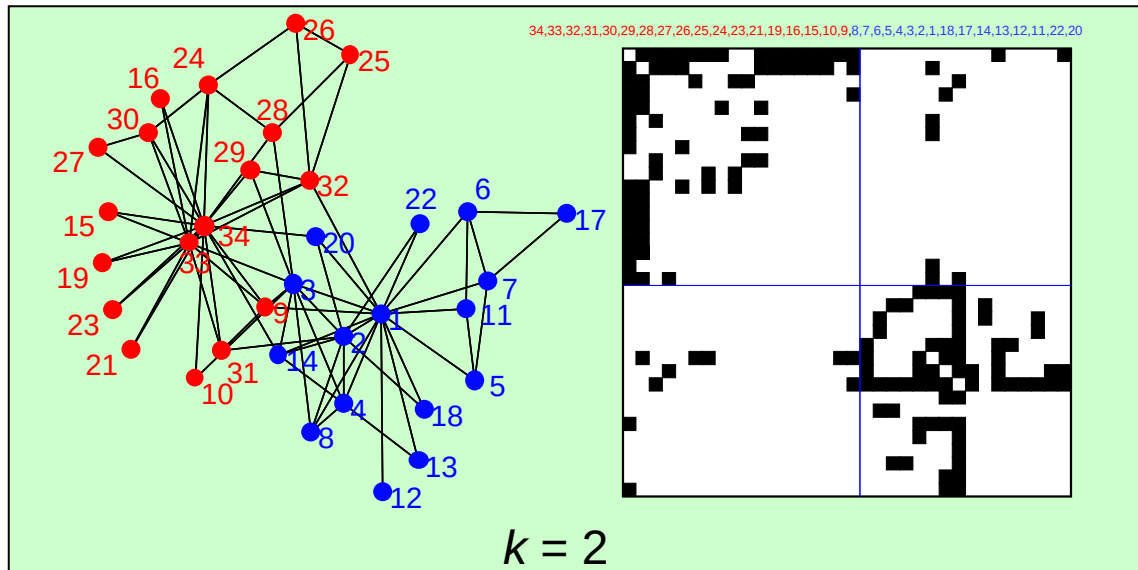
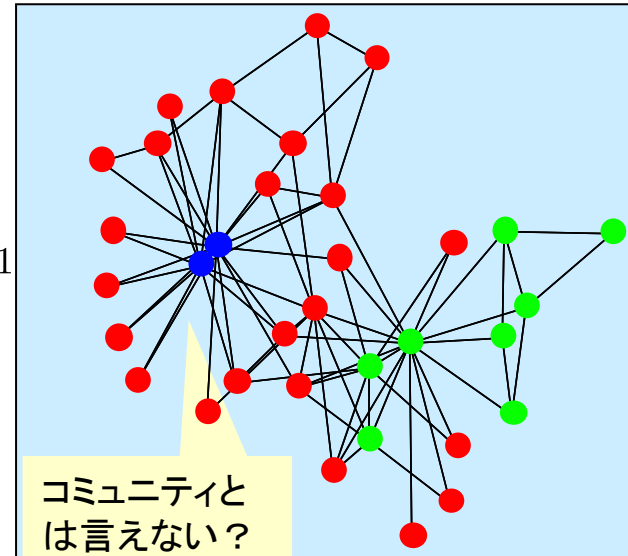
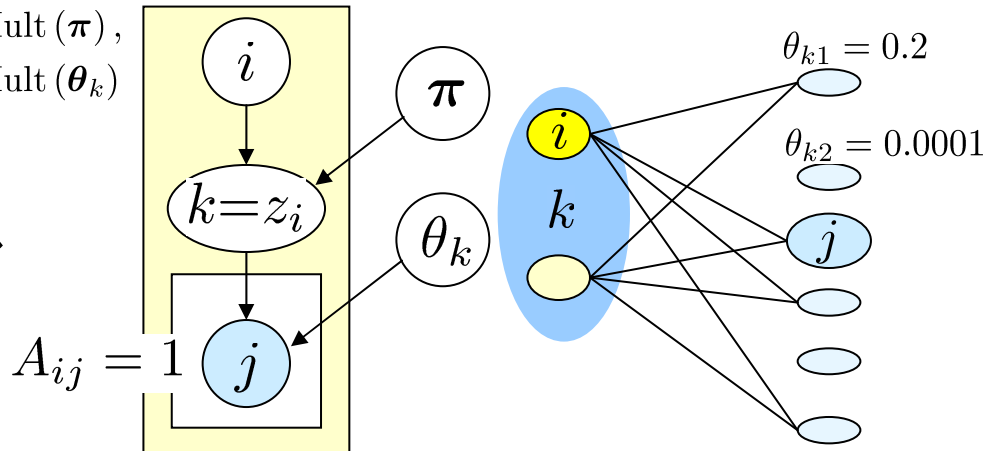
同じクラスのノードは、同じ相手にリンクを張る

$$k = z_i \sim \text{Mult}(\pi),$$

$$j \sim \text{Mult}(\theta_k)$$

$$A_{ij} = 1$$

★prior なし



Assortative vs disassortative

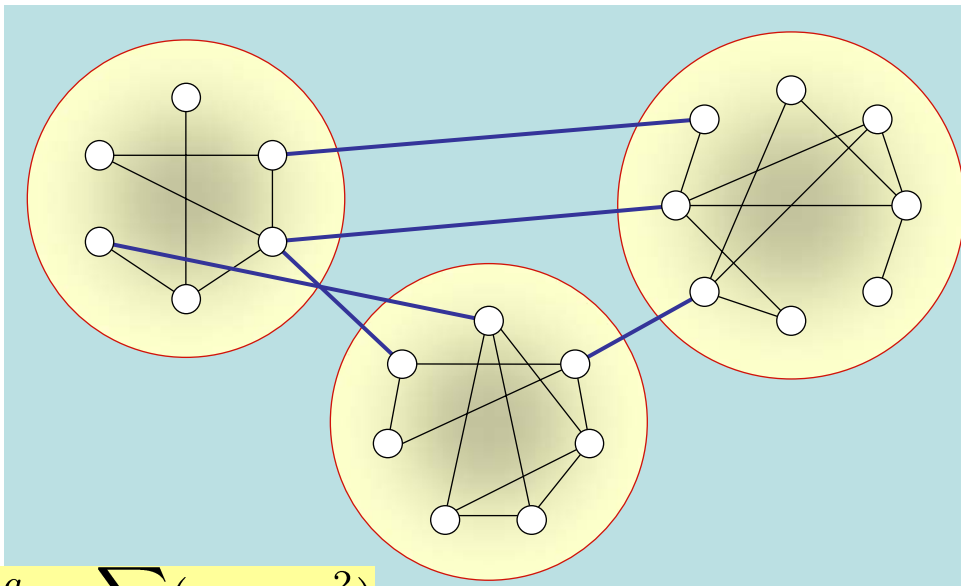
assortative:

- コミュニティ内のリンクが密、コミュニティ間のリンクが疎

disassortative:

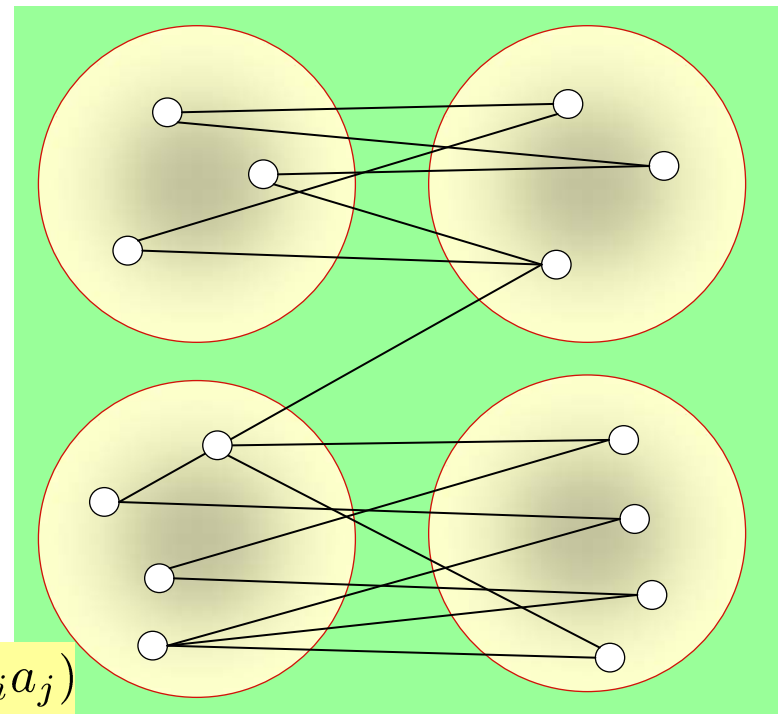
- コミュニティ間のリンクが密、コミュニティ内のリンクが疎

assortative



$$Q^a = \sum_i (e_{ii} - a_i^2)$$

disassortative



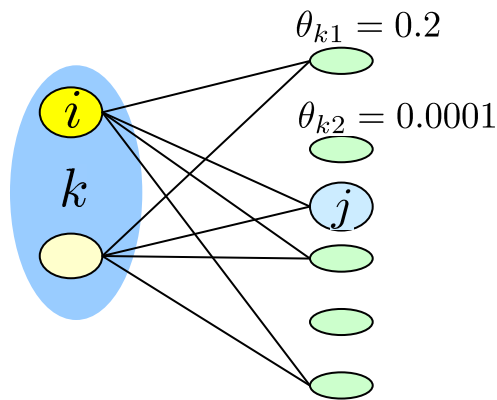
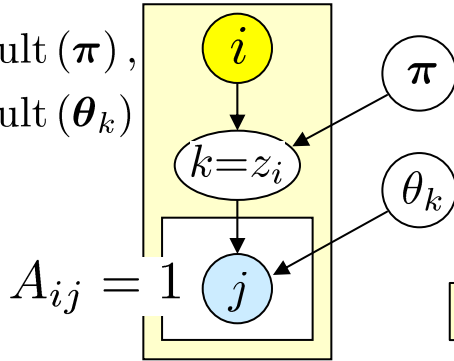
$$Q^d = \sum_{i>j} (e_{ij} - a_i a_j)$$

DPに基づく「無限」クラスタリング

Mixture Model の「無限」化

Mixture Model

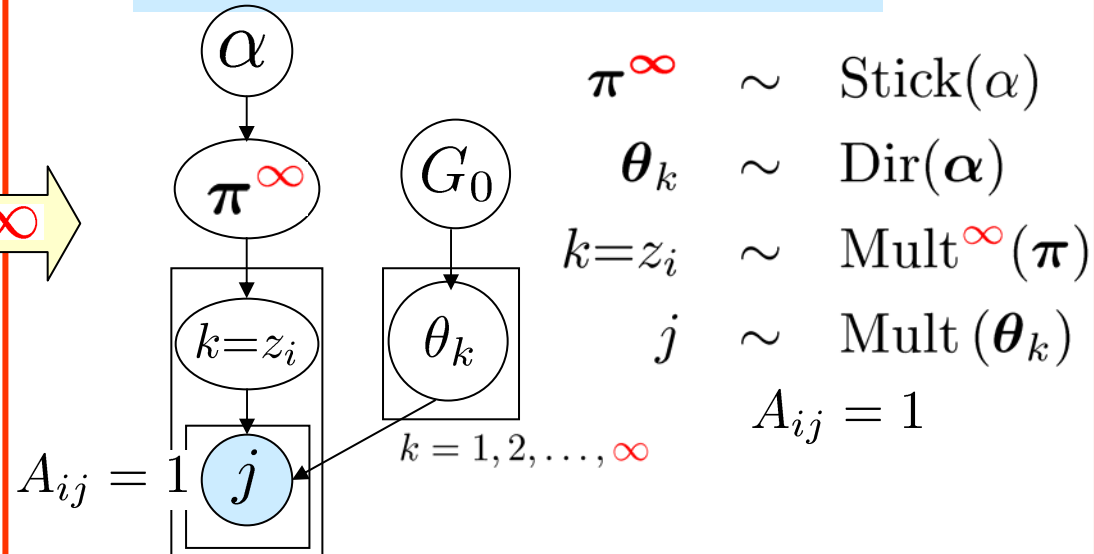
$$\begin{aligned} k = z_i &\sim \text{Mult}(\pi), \\ j &\sim \text{Mult}(\theta_k) \\ A_{ij} &= 1 \end{aligned}$$



Chinese Restaurant Process

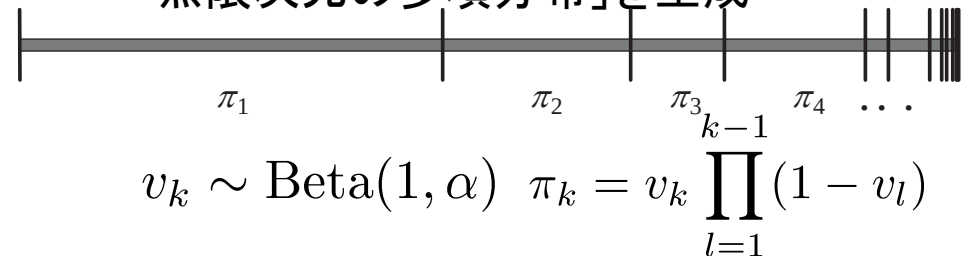
$$Z|\alpha \sim \text{CRP}(\alpha)$$

Infinite Mixture Model



Stick Breaking Process

= 「無限次元の多項分布」を生成



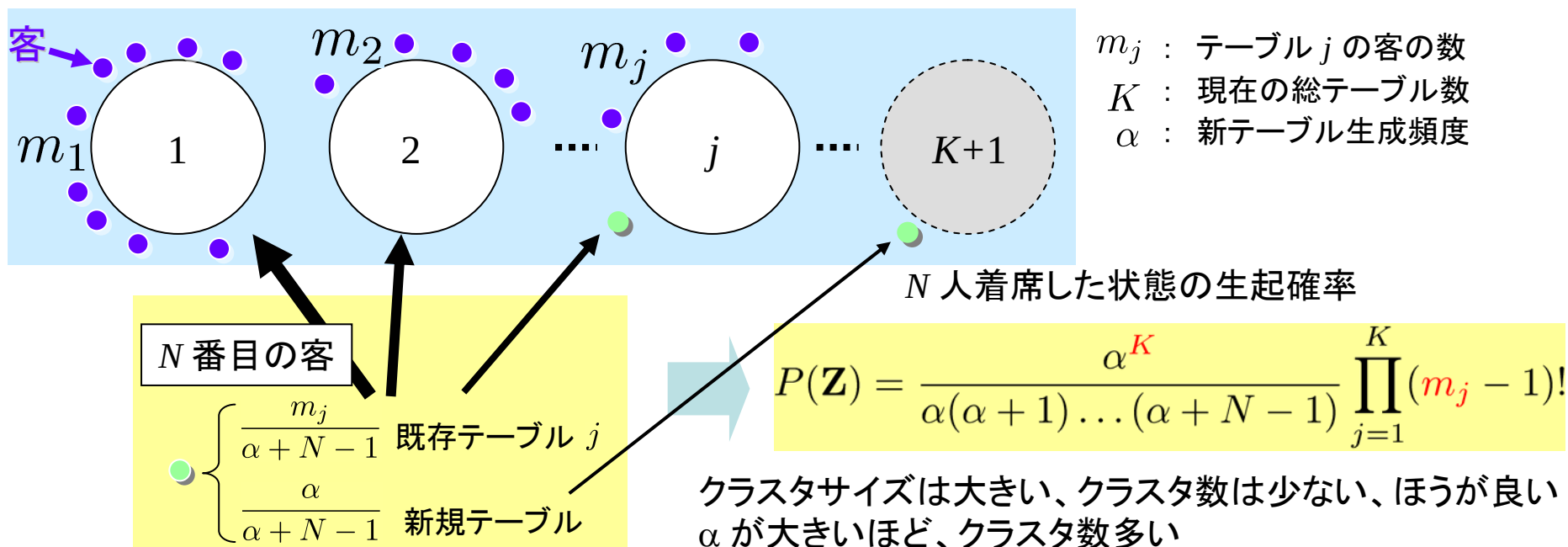
Chinese Restaurant Process

クラスタ数をあらかじめ固定せず、データごとに最適なクラスタリングをしたい

「自然な」「無限個クラスタ」のモデル

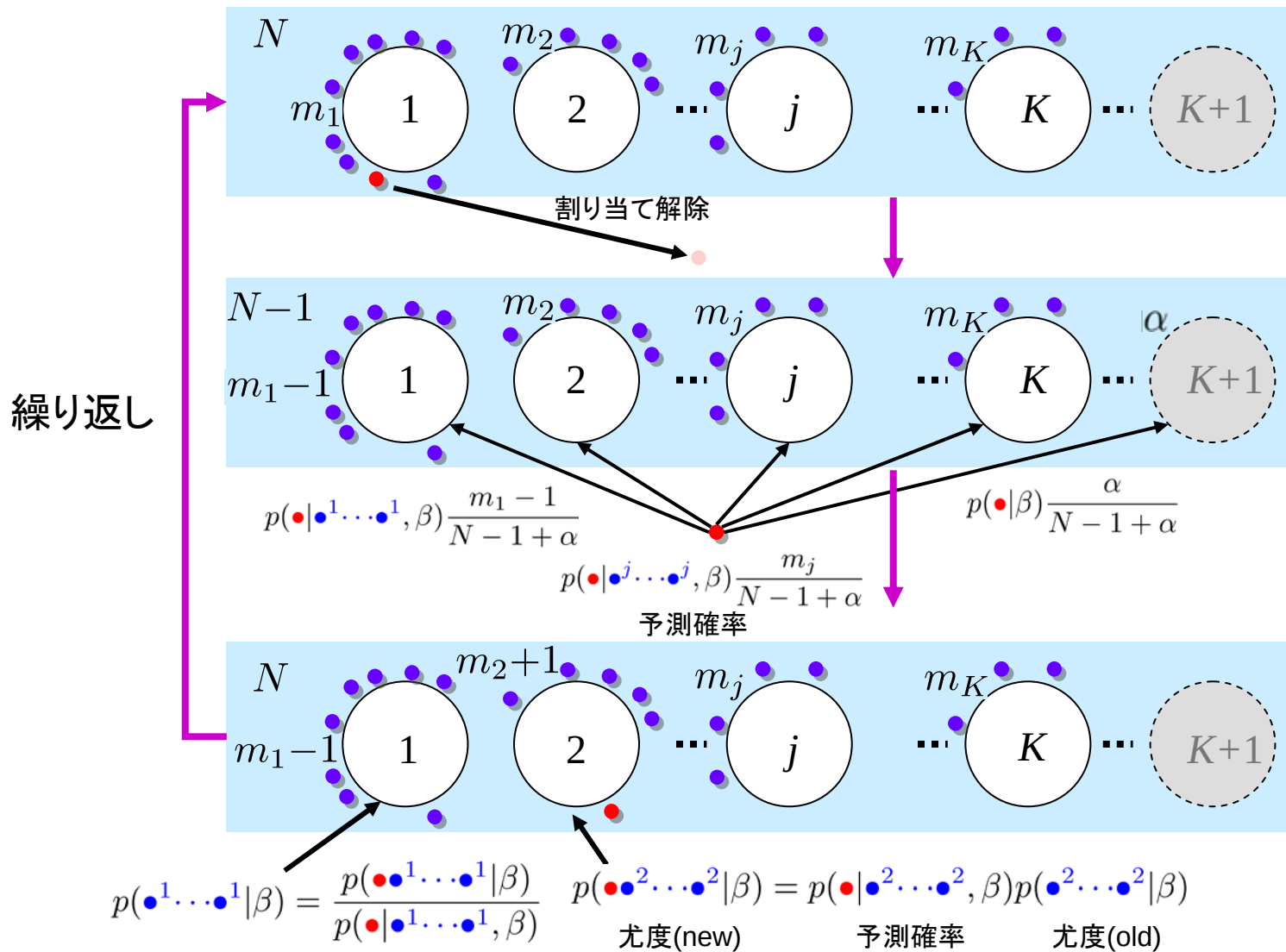
Chinese Restaurant Process

- すでにたくさん人がいるテーブル(クラスタ)ほど新たな客が座りやすい
- 一定の確率で新しいテーブルに座る場合もある



- 優先選択モデル(Preferential Attachment)と共通する考え方
- 中立進化説(自然淘汰が存在しない場合の進化)のモデル

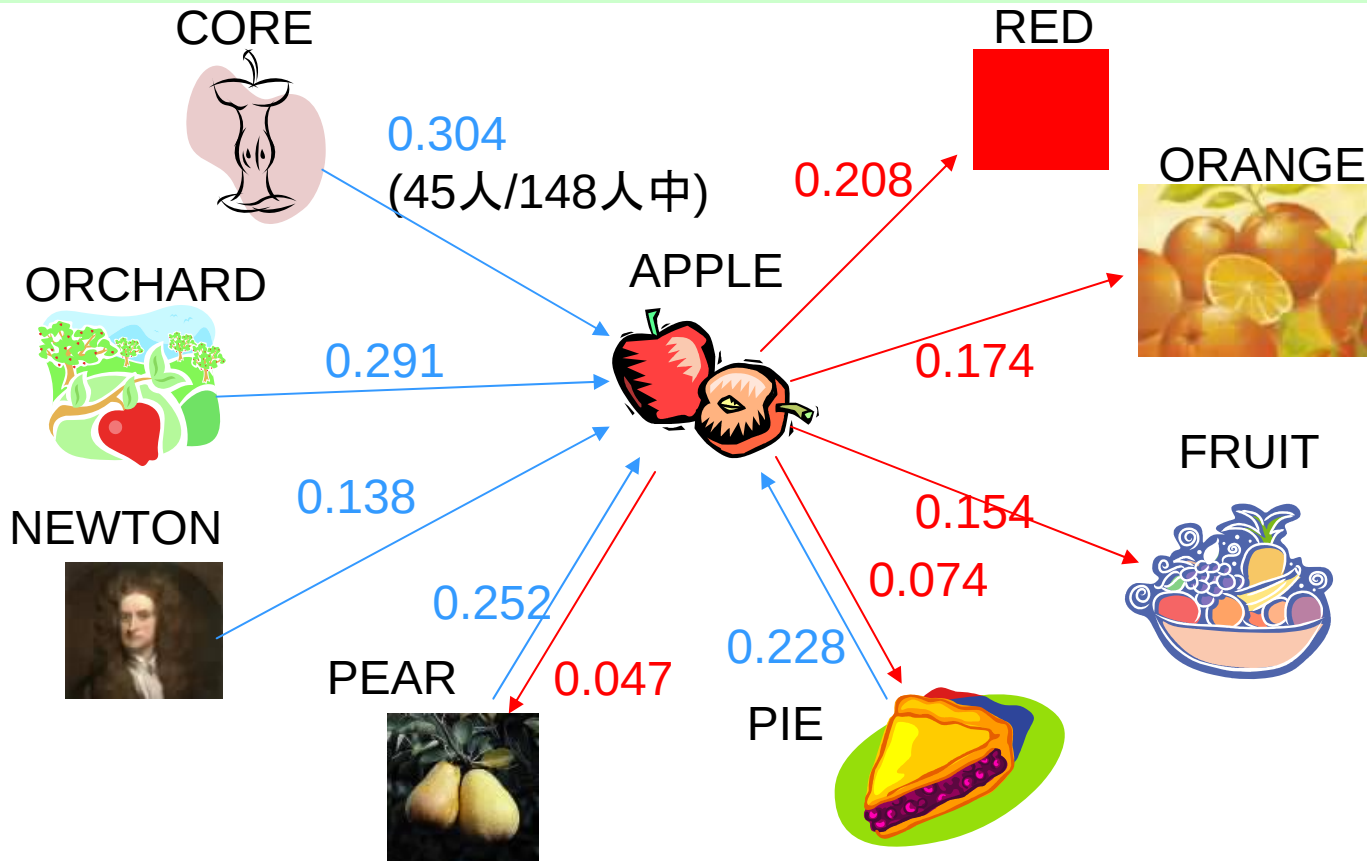
Collapsed Gibbs Sampling



$$P(\mathbf{Z}) = \frac{\alpha^K}{\alpha(\alpha + 1) \dots (\alpha + N - 1)} \prod_{j=1}^K (m_j - 1)!$$

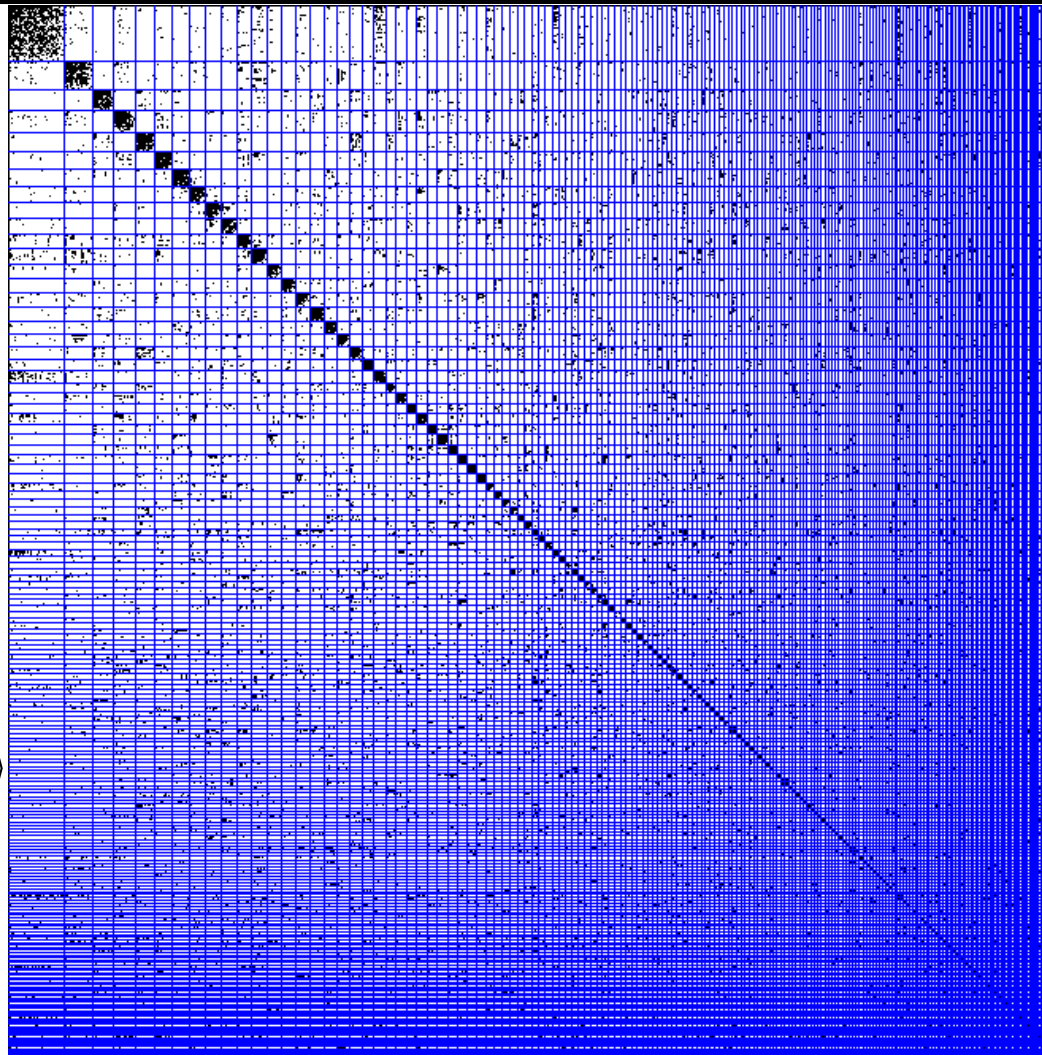
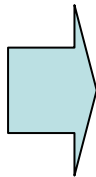
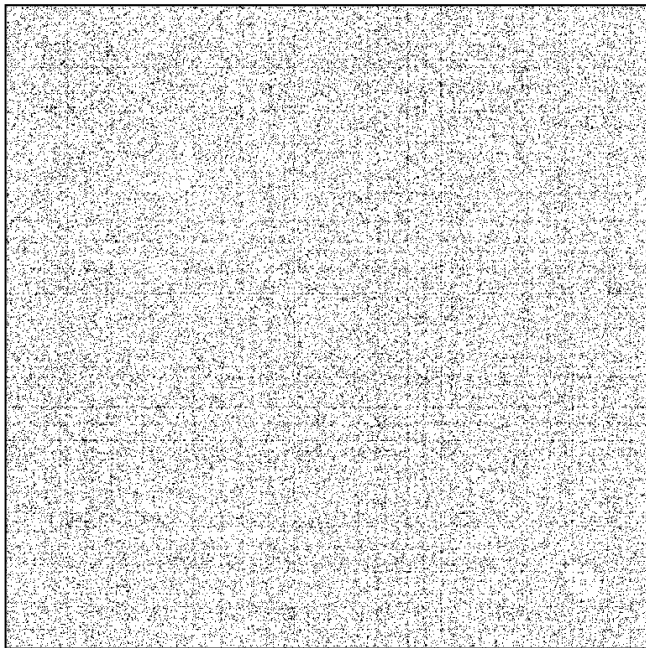
単語連想ネットワーク

元データは、南フロリダ大学心理学科のプロジェクトで作成
被験者数：平均149人(±15程度)
被験者にキューとなる単語を与え、関連する単語を連想してもらう



双方向のリンクの強さの和を一定の閾値(0.025)で無向化
⇒ ノード数：7,207、リンク数：31,784

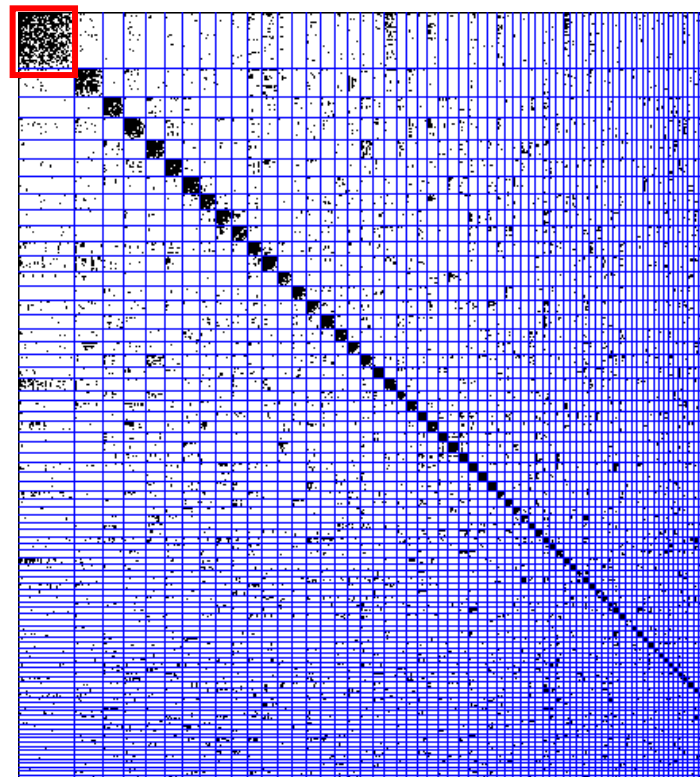
単語連想ネットワークのクラスタリング例



単語連想ネットワークの例

FOOD	PASTRY	FROST	CABBAGE	UTENSIL	INPUT
VINEGAR	CATCHER	BITTER	COLESLAW	LEMONADE	OUTPUT
TASTE	PITCHER	SWEET	CORNBEEF	HASH	OUTTAKE
TOPPING	RACKET	TANGY	LETTUCE	JELLO	ISAAC
HAIRSPRAY	PEEL	TART	CUPBOARD	CRUMBS	KERNEL
LUNCH	BANQUET	OLIVE	CAFE	HUNGER	TEQUILA
VEGETABLES	DINNER	STALLION	DINER	SOGGY	SUCKER
CHIPS	EAT	BLACKBERRY	CAFETERIA	NESTLES	LOW_FAT
COOKIES	FEAST	JAM	DESSERT	CUCUMBER	SAP
FRESH	CRUNCH	WINE_COOLER	FROSTED	PICKLE	PARKAY
COCKTAIL	SALAD	STARCH	FROSTING	TOMATO	MARTINI
GRAIN	SALOON	MILKSHAKE	ICE_CREAM	HOTNESS	MASHED
BOWL	CHARCOAL	PANCAKES	ICING	MUGS	PENICILLIN
BUTTER	HAMBURGER	CHECKERS	PUDDING	CUSTARD	PLANTERS
CANDY	SAUCE	BOILED	SNACK	MUSTARD	VITAMIN_C
CHOCOLATE	BARLEY	CHEESE	SPATULA	DADDY	POD
SOUP	HERB	OSCAR_MAYER	CAKES	DANISH	PILGRIMS
PIE	HOPS	SANDWICH	CAMPBELL	FAMINE	SPOONS
CRACKERS	OATS	RESTAURANT	SARDINE	PARSLEY	POPS
FEED	SPICE	BOTTOMLESS	CRUNCHY	DELICIOUS	TOMATOES
PICNIC	WHEAT	CEREAL	SCAR	SCUMPTIOUS	SHRIVEL
APPLE	CRACK	BOON	TREAT	SAVOR	SAVOR
BAKED	PICKLES	HOT_DOG	PRETZEL	RICE-A-RONI	RUBBERMAID
BANANA	CHERRY	COBBLER	PEANUT BUTTER	STUFFING	SNICKERS
FRUIT	GRAPE	JUICE	CRACKLE	TAPIOCA	THINS
ORANGE	FRANKS	ORCHARD	CHEDDAR	FORK	WHIPPED_CREAM
PEACH	PEAR	PLUM	CREAM	HUNGRY	TUMMY
APPLE_JUICE	ORANGE_JUICE	PLATE	GOURMET	STARVE	
TRAY	BREAKFAST	CRISP	CHERISH	NOG	
FRY	SOUR	STALE	FLAVOR	SCRAMBLE	
TEA	BAGEL	BREAD	CHOMP	SCRAMBLED	
CREAM_CHEESE	DOUGHNUT	BAKE	MUNCH	ESCARGOT	
BROIL	CAKE	COOKIE	CHEWY	ETHIOPIA	
FLOUR	COOKED	FRIED	NUGGETS	FAMISHED	
POTATO	POTATOES	BAKER	PEPPER	OPTICS	
BAKERY	BUTCHER	CHEF	SPICY	FLAPJACKS	
			SILVERWARE	FRIES	
			SLICE	MEATBALLS	
			DORITOS	OREGANO	
			PRETZELS	PRODUCE	
			MOUSSE	SERVER	
			VANILLA	FORKS	
			YOGURT	KETCHUP	
			YUMMY	PRUNE	
			CHOWDER	TANGERINE	
			HUNK	GEL	
			CITRUS	GEORGIA	
			DRESSING	PASTE	
			SEASONING	SHRED	
			WAITRESS	GREASY	
			STIR	LIME	
			TEAPOT	GREENS	
			POTATO_SALAD	GRITS	
			CONDIMENT	HOST	
			CONSUME	PUMPKIN	
			COOKBOOK	HAWAIIAN	
			RECIPE	HERBAL	
			SKILLET	SYRUP	
			FIG	IDAHO	
			MOLASSES		

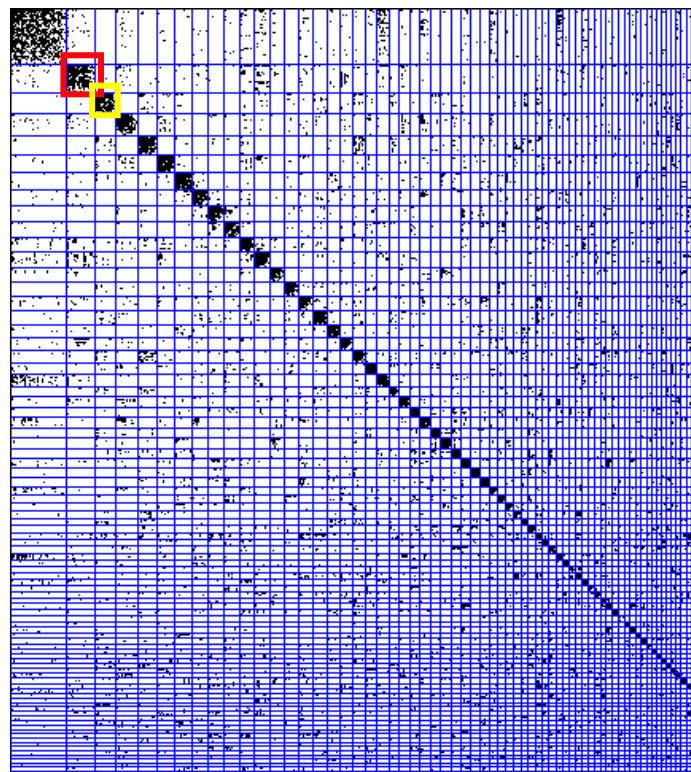
食へ物



単語連想ネットワークの例

OVERFLOW	TIDE	
ABUNDANT	WAVES	
SPILL	MERMAID	
AEROSOL	WATERFALL	
BOAT	DAM	
SAILOR	DEEP	
SHIP	LAGOON	DOLPHIN
WATER	DIVE	FIN
GLIDE	DIVING	FLIPPER
NAVIGATOR	CANAL	TUNA
LAKE	CANOE	UPSTREAM
POND	DECK	DOWNSTREAM
POOL	FISHING	FLOW
ALLIGATOR	MOAT	SALMON
GULLY	OAR	FAUCET
REPTILE	PADDLE	DRENCH
ANCHOR	PORT	POUR
LIZARD	RAFT	MALLARD
MARINE	ROW	SLAM
OTTER	SAIL	EXCESS
PORPOISE	SNAIL	SPIGOT
OCEAN	SILVER	FERN
BAY	SHOAL	REEL
SURFACE	SHOAL	FINN
TANK	SWIMMING	FIVE_0
SWIMMER	MARSH	FLOATER
ATLANTIC	BON_VOYAGE	SLIMY
DRIFT	BOULDER	GUSH
BABBLING	LOAF	OLD_FAITHFUL
BROOK	UNDERWATER	SPOUT
BAHAMAS	WINDY	YELLOWSTONE
CRUISE	LEAK	GOO-GAA
SNORKEL	CREEK	SNAGGED
BAIT	STREAM	KEY WEST
FISH	DUNE	MOISTEN
SHRIMP	BUOYANT	SPLASH
CLAM	C.B.	PUDDLE
FLOAT	ERIE	SEA_LION
RIVER	PANAMA	OVERFILL
OYSTER	SWIM	PEBBLES
BUCKET	VALLEY	PIPES
REEF	WET	PLUNGER
SINK	MUSSEL	RAIN_FOREST
DOCK	SCALLOP	SPA
HARBOR	SEAFOOD	WASHBASIN
INLET	DRAIN	WEIGH_T
SEA	DRANO	SPLISH
SHORE	UNCLOG	SPRITZ
BEACH	DRIP	SQUIRT
COAST	CROAK	SUBMERGE
CORAL	HOOVER	
GULL	DEW	
PELICAN	DROWN	
PIER	DEPTH	
ROCKS	SHALLOW	
SAND	FLOOD	
SEAGULL	DISNEY_WORLD	
SEASHORE	SCUBA	
SHELL	DIVER	
SURF	SWIMMING	

海



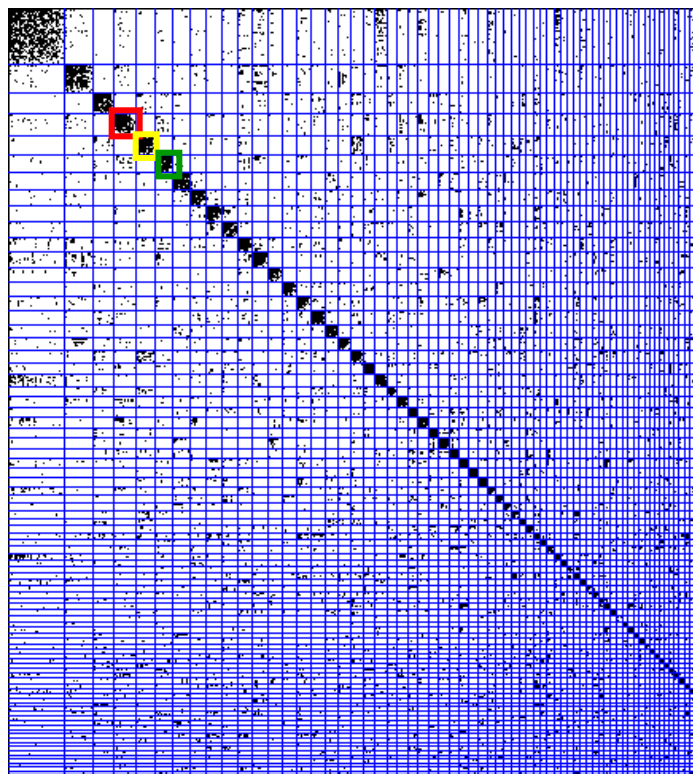
ALPHABET	READ	
AN	TEXT	
LETTER	WORDS	
TYPE	WRITING	
ACCENTUATE	ESSAY	
WRITER	SPEECH	PHRASES
BILLBOARD	CARDBOARD	POE
PAPER	OUTLINE	TYPING
BOOK	BRITANNICA	SCRABBLE
MAIL	COLUMN	SENTENCES
LIST	CAPTION	SPELLING
DOCUMENT	HEADING	SYNONYM
TERMS	REPORT	WORD_PROCESSOR
ALMANAC	DIAGRAM	
DICTIONARY	GRAFT	
ENCYCLOPEDIA	MARK	
FACTS	SIGNATURE	
FARMERS	CLEARING_HOUSE	
INFORMATION	MARKER	
ALOUD	SENTENCE	
READING	PRINTING	
LETTERS	SOFTWARE	
STATEMENT	TYPEWRITER	
ANTITHESIS	WHIZ	
PAD	CONS	
STAMP	PROSE	
AUTHOR	PENCIL	
EDITOR	CURSIVE	
POET	ISSUE	
PUBLISHER	TERM	
TITLE	LEXICON	
WRITE	INGEST	
BACKPACK	READERS	
BOOKBAG	TOPIC	
PARCEL	DISSERTATION	
TOTE	DOODLE	
SONNET	EMPHASIZE	
SERIES	PARAGRAPH	
POEM	PHRASE	
POETRY	ENVELOPE	
BEDTIME	QUILL	
SCROLL	STAPLER	
VERSE	TOWARD	
BINDER	SPOTLIGHT	
FOLDER	UNDERLINE	
NOTEBOOK	HIGHLIGHTER	
DOT	PAPERCLIP	
RAP	HOLD_BACK	
INK	THEME	
PEN	JOT	
BLOTTER	POSTMAN	
BULLETIN	POST_OFFICE	
BRITANNICA	PERMANENT	
JOURNAL	MEMO	
LITERATURE	MISSPELL	
NOTE	SPIRAL	
NOTES		
NOVEL		
PAGE		

文字

単語連想ネットワークの例

AARDVARK	WOLF	LEE
ANIMAL	ZOO	LOUISIANA
ANTEATER	ANIMALS	MANATEE
PINK_PANTHER	HUNTING	MINNESOTA
ADAMS	PETS	PEEVE
PUPPY	ANTLER	POLKADOTS
SAFARI	CANTELOUPE	SHREW
AFRICA	ANTLERS	SOFTENER
GAZELLE	HORNS	STAINS
JUNGLE	GORILLA	TOCK
LION	HAIRY	VIETNAM
ZEBRA	TARZAN	
PENGUIN	BAMBI	
CAT	SMOKY	
CROCODILE	BRAWL	
STAG	BARK	
FURRY	MEOW	
FURY	GLOVE	
FIERCE	CLEAVER	
ANTELOPE	WHALE	
APE	FAWN	
BEAR	PIGEON	
BEAVER	MOUSE	
BUCK	SAFARI	
CAGE	RUN	
CHIMPANZEE	COUGAR	
CLAW	LICE	
COUGAR	RODENT	
COYOTE	ROCKY	
DEER	KITTEN	
DOG	MOUSE	
DONKEY	CAVE_MAN	
ELEPHANT	CHEETAH	
ELK	LAMB	
EWE	TALON	
FOX	CLEAT	
FUR	COBRA	
GIRAFFE	CONNIVING	
GOAT	JOHN	
HUNT	CUDDLY	
LEOPARD	DALMATIAN	
MAMMAL	DOCILE	
MINK	QUACK	
MONKEY	JANE	
MOOSE	RAE	
MULE	KONG	
PANTHER	TUSK	
PAW	PUFFY	
PET	NAVEL	
PREDATOR	WUZZY	
PRIMATE	GERMAN	
RACCOON	IN_THE_MIST	
RAT	IVY	
SHEEP	HAGAAR	
SPHINX		
TAIL		
TAME		
TIGER		
TRAP		
VET		
WALRUS		

動物



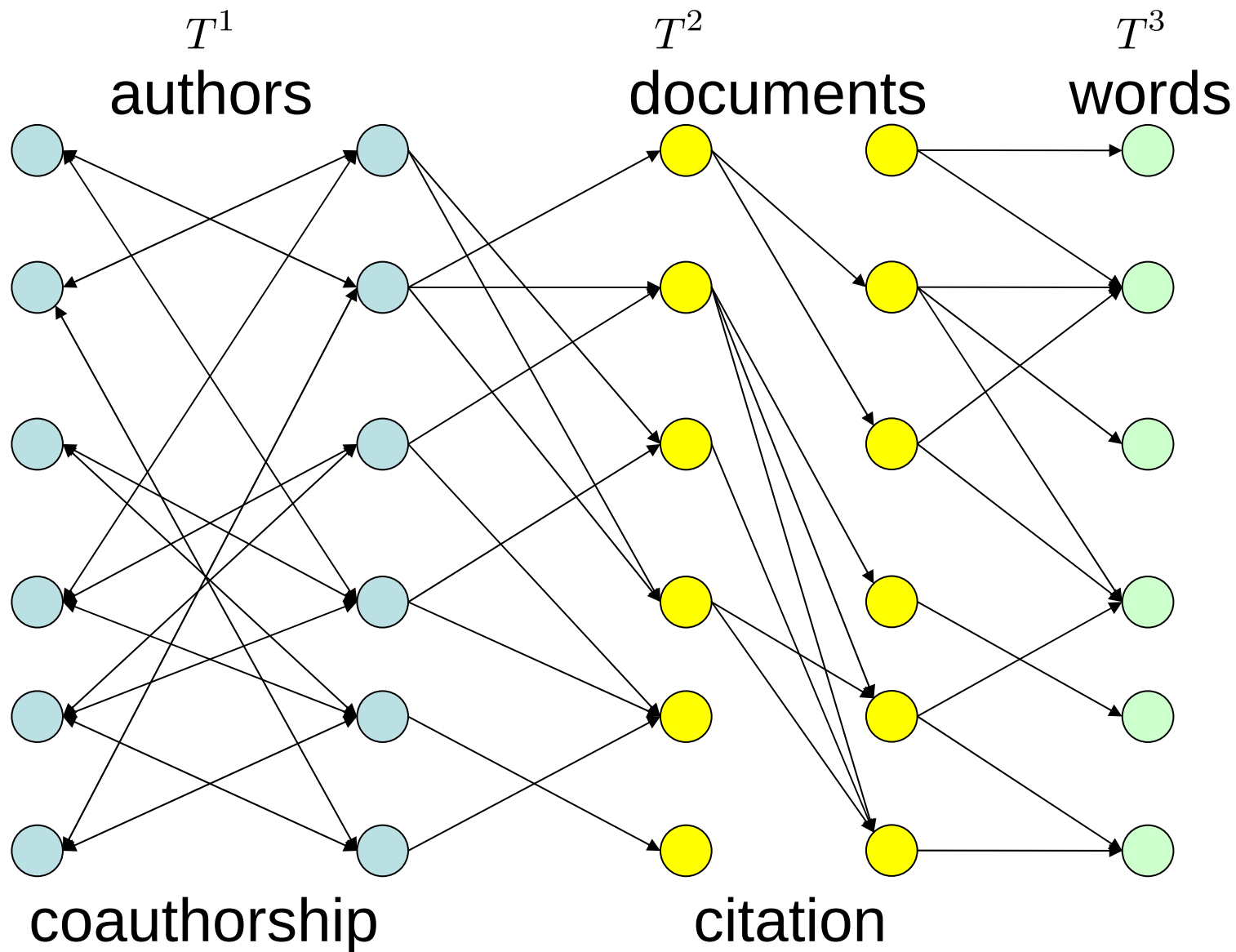
20/20	WISE
HINDSIGHT	IDEA
UNDERSTAND	CONCEPT
KNOW	SUBCONSCIOUS
NOTICE	KNOWING
REALIZE	CONTEMPLATE
SEE	PONDER
KNOWLEDGE	RECONSIDER
OBJECTIVE	IMAGINATION
SUBJECTIVE	IMAGINE
OPINION	INVENT
WISDOM	IMAGINATIVE
FORESIGHT	INGENIOUS
EINSTEIN	INVENTOR
AMOeba	CYNIC
DATA	WHAT
THINK	PERCEPTION
ANALYZE	REVEAL
COMPARE	HOW
FIGURE	ESP
INTERPRET	INSIGHT
LOOK	PREDICT
DETERMINE	FAITH
GUESSES	HYPOTHESIS
THESIS	THEORY
RATIONALE	PLAN
ARISOTLE	INJUDIC
ASSUME	PSYCHIC
PRESUME	FRAME
SUPPOSE	GAZE
CONCENTRATE	THEOREM
FOCUS	TALENTED
VISION	WHO
VIEW	WHERE
REASON	WHEN
SO	I.Q.
WHY	IMAGE
PHILOSOPHY	SURVIVAL
BELL_TOLLS	INTELLECTUAL
MIND	IQ
STARE	LIGHT_BULB
GLARE	PERISCOPE
SUBJECT	PERCEIVE
BRAIN	LAPSE
INTELLECT	REFLECT
INTELLIGENCE	OBSERVE
LOGIC	ORGANISM
SMART	WATCHED
THINKING	WHOO
THOUGHT	PLATO
BRILLIANT	SOCRATES
GIFTED	THINKER
INTELLIGENT	POSTULATE
GENIUS	PYTHAGOREAN
LENS	THEY
DISCOVER	
CARED	
DECIDE	
DECISION	
CLASSY	

知能

DRESS	WARDROBE
CLOTHES	FABRIC
SLEEVE	SATIN
SUIT	SILK
ASSEMBLE	FAD
CLOSET	FASHION
ATTIRE	LACE
CASUAL	OUTFIT
FORMAL	SEAM
WEAR	STITCH
COTTON	VEST
CROCHET	LAPEL
FUZZ	WOOL
YARN	NYLON
BANDANNA	CREASE
CLOTH	KNIT
SCARF	CUFF_LINK
ROBE	PROM
PANTS	DECORATIVE
SASH	DENIM
TROUSERS	HEM
RACK	SLACKS
TUXEDO	DUNNAPES
VELVET	FEELING
KNITTING	FIREMAN
QUILT	FLANNEL
BLEACH	FLUFF
JEANS	TUX
BLOUSE	RIP
SHIRT	RUFFLE
SKIRT	PANTY
SWEATER	LEVI'S
COLLAR	KAT
RIBBON	SEWING
TIE	LOOM
BOOBS	MATCHING
BRA	THREAD
SHELF	NIGHTIE
SUEDE	SEVENTIES
SHORTS	SIXTIES
UNDERWEAR	SETTER
HOOD	STUFFER
PANTIES	TAKE_OFF
DRAWER	
COAT	
JACKET	
MUTTON	
SEW	
ZIPPER	
CABINET	
CABOODLE	
MATERIAL	
GOWN	
CAPONE	
DRESSY	
INFORMAL	
POLYESTER	
PLAID	
CHECKERED	
PATTERN	
STOCKING	

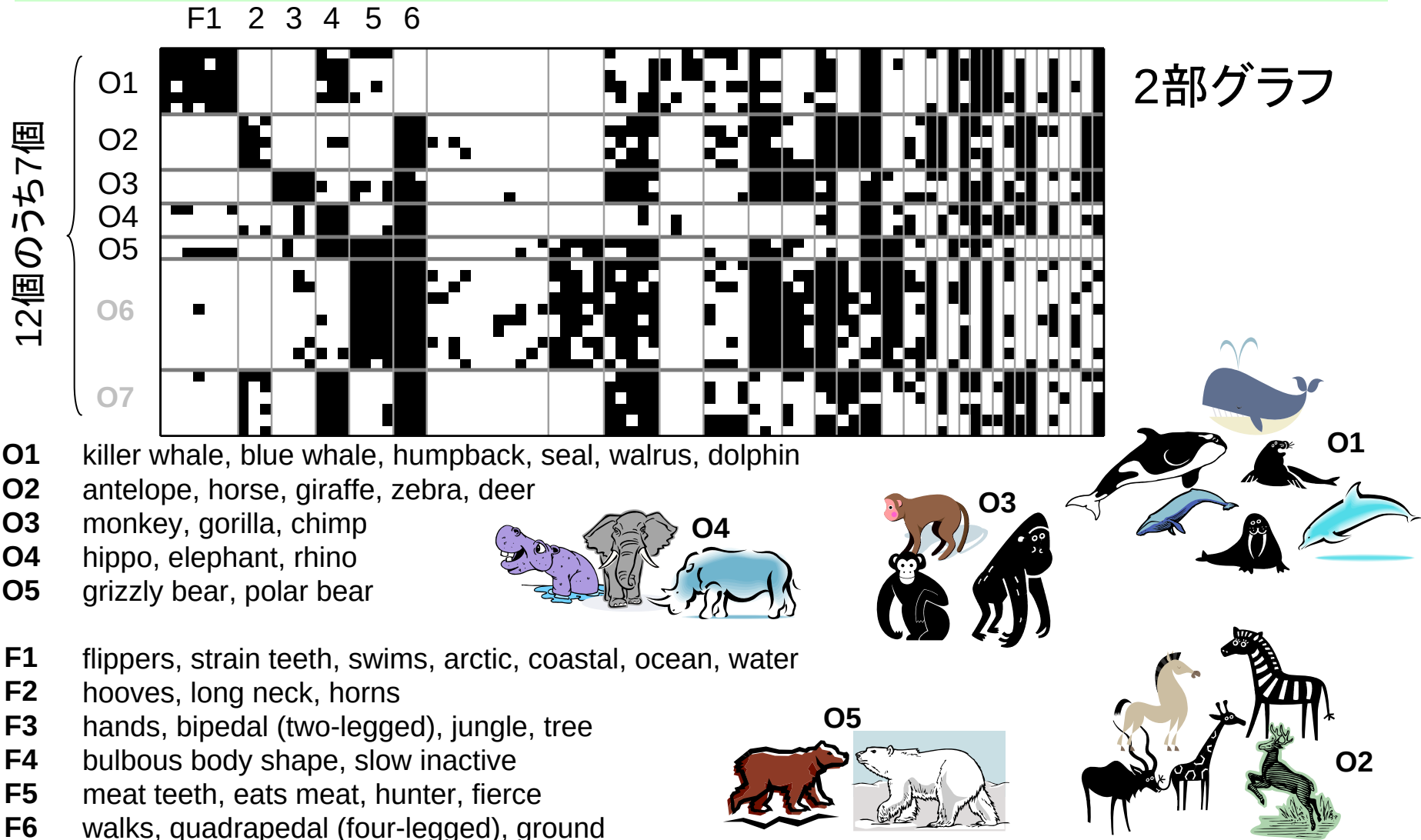
服飾

多ドメイン間の多重関係モデル



ブロックモデルの例：動物とその特徴

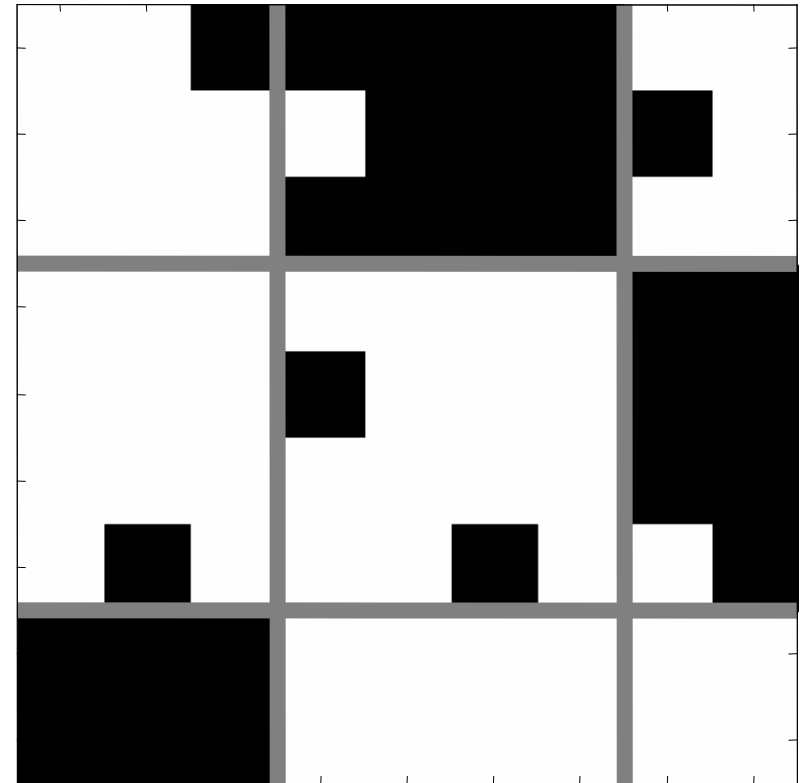
50種類の動物と85の特徴からなるデータをクラスタリングした例



ブロックモデルにおける関係生成モデル(1)

左のリンク確率にしたがって右のようにリンクが生成。

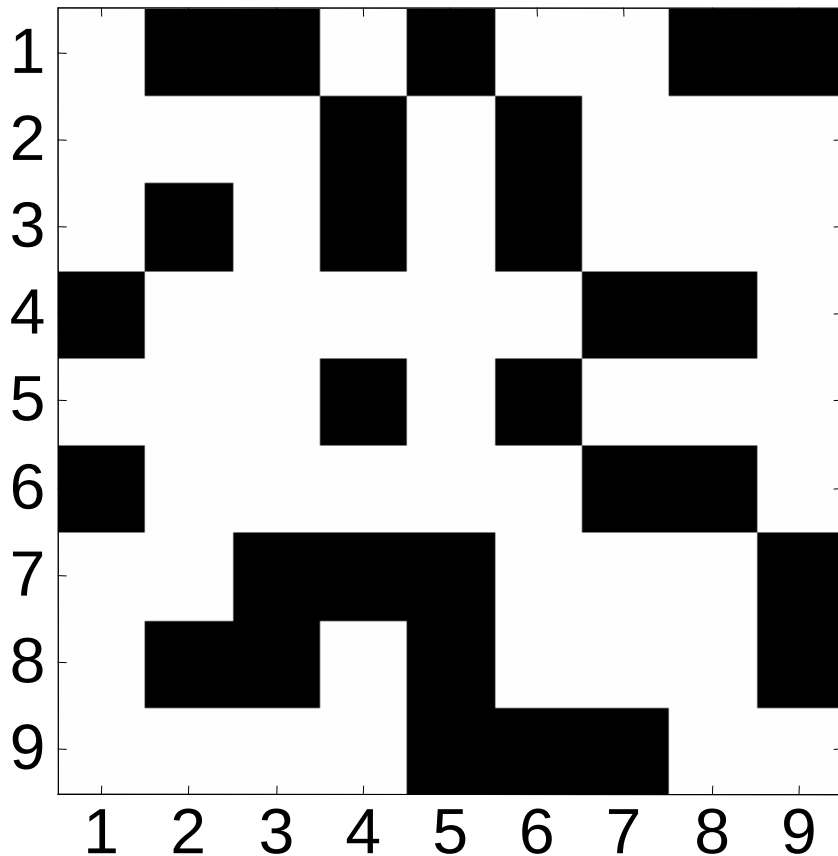
$\eta_{11} = 0.1$	0.9	0.15
0.05	0.1	0.85
0.98	0.01	0.01



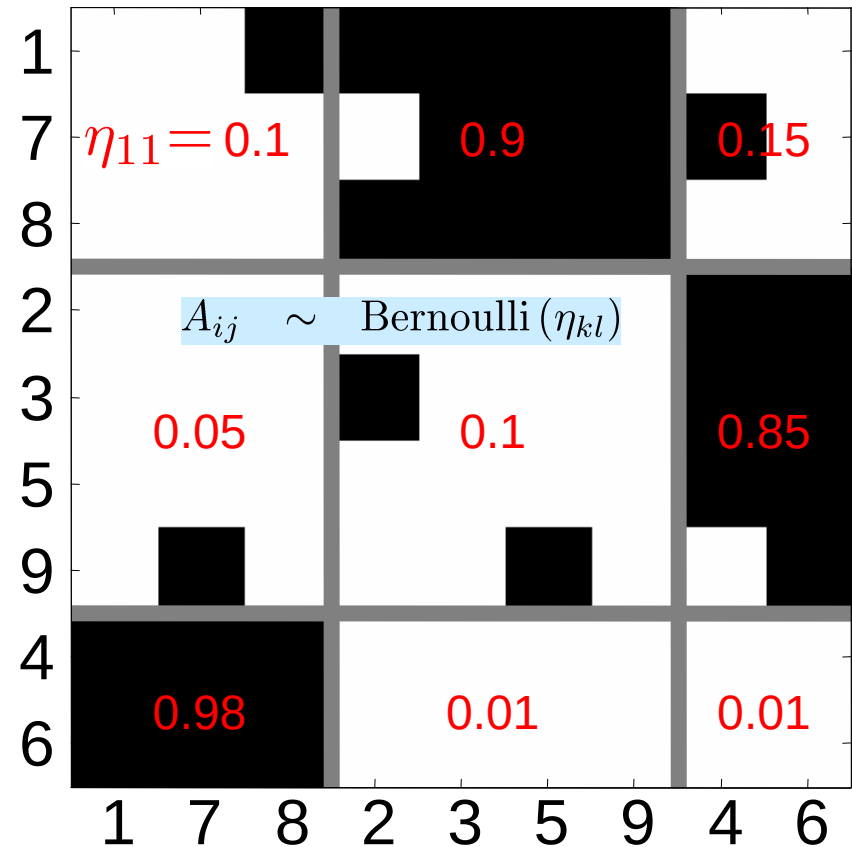
$$A_{ij} \sim \text{Bernoulli}(\eta_{kl})$$

ブロックモデルにおける関係生成モデル(2)

実際に観測される接続行列は、左のようにランダムにシャッフルされている



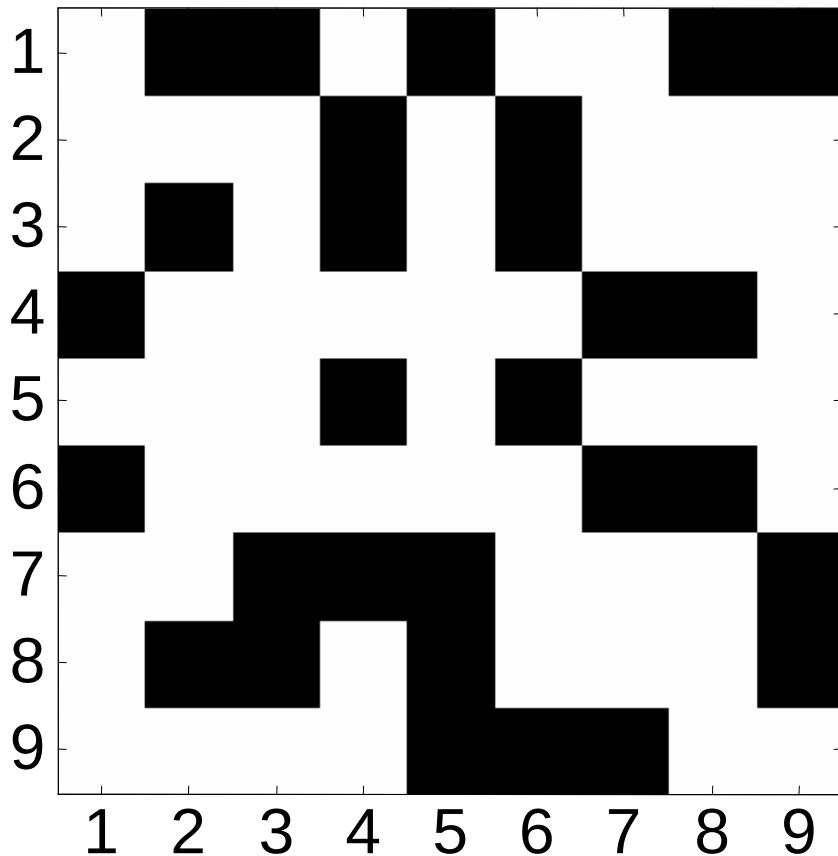
リンクのあるところが黒、ないところが白



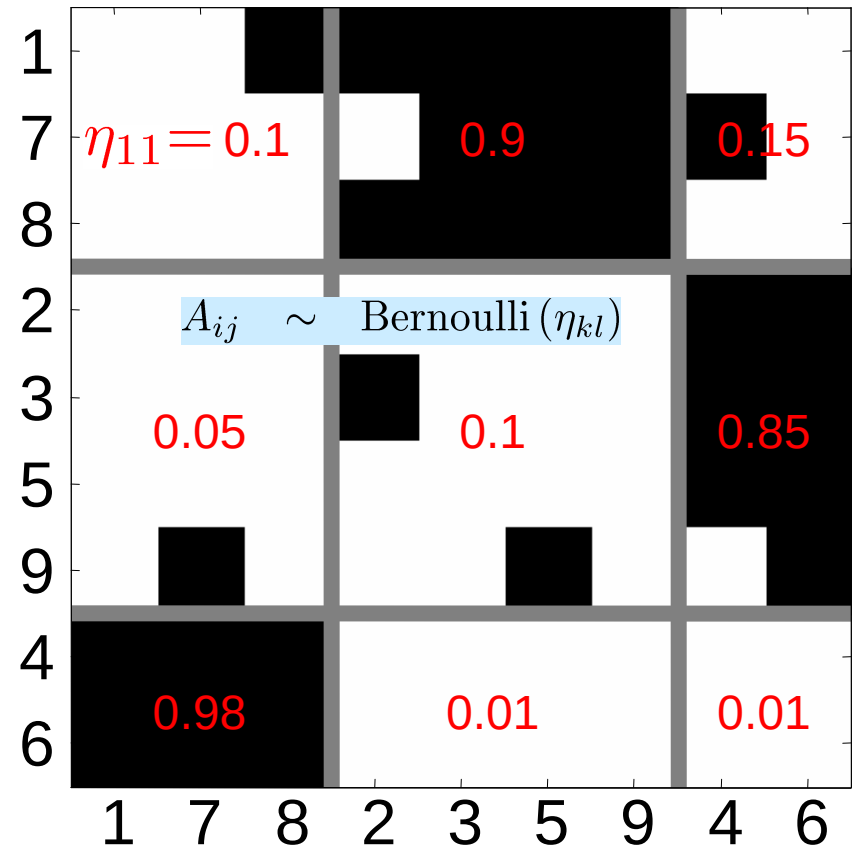
同一クラス内は、白・黒がはっきり分かれる

ブロックモデルにおけるクラスタリング

左の「接続行列」をうまく並び替えて、右のような、隠れた「ブロック構造」を見つけることに相当する



リンクのあるところが黒、ないところが白



同一クラスタ内は、白・黒がはっきり分かれる

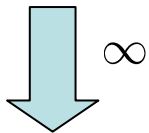
Stochastic Block Models

リンクを張る相手もクラスタリングされている

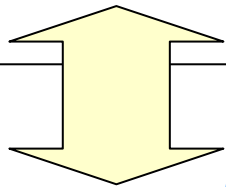
Estimation and prediction for stochastic blockstructures
K. Nowicki and T.A.B. Snijders, JASA 96 (455), 2001

Stochastic block models

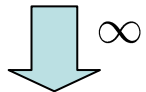
同じクラスのノードは、みな
同じ相手クラスにリンクを
{張る、張らない}



Infinite Relational Model

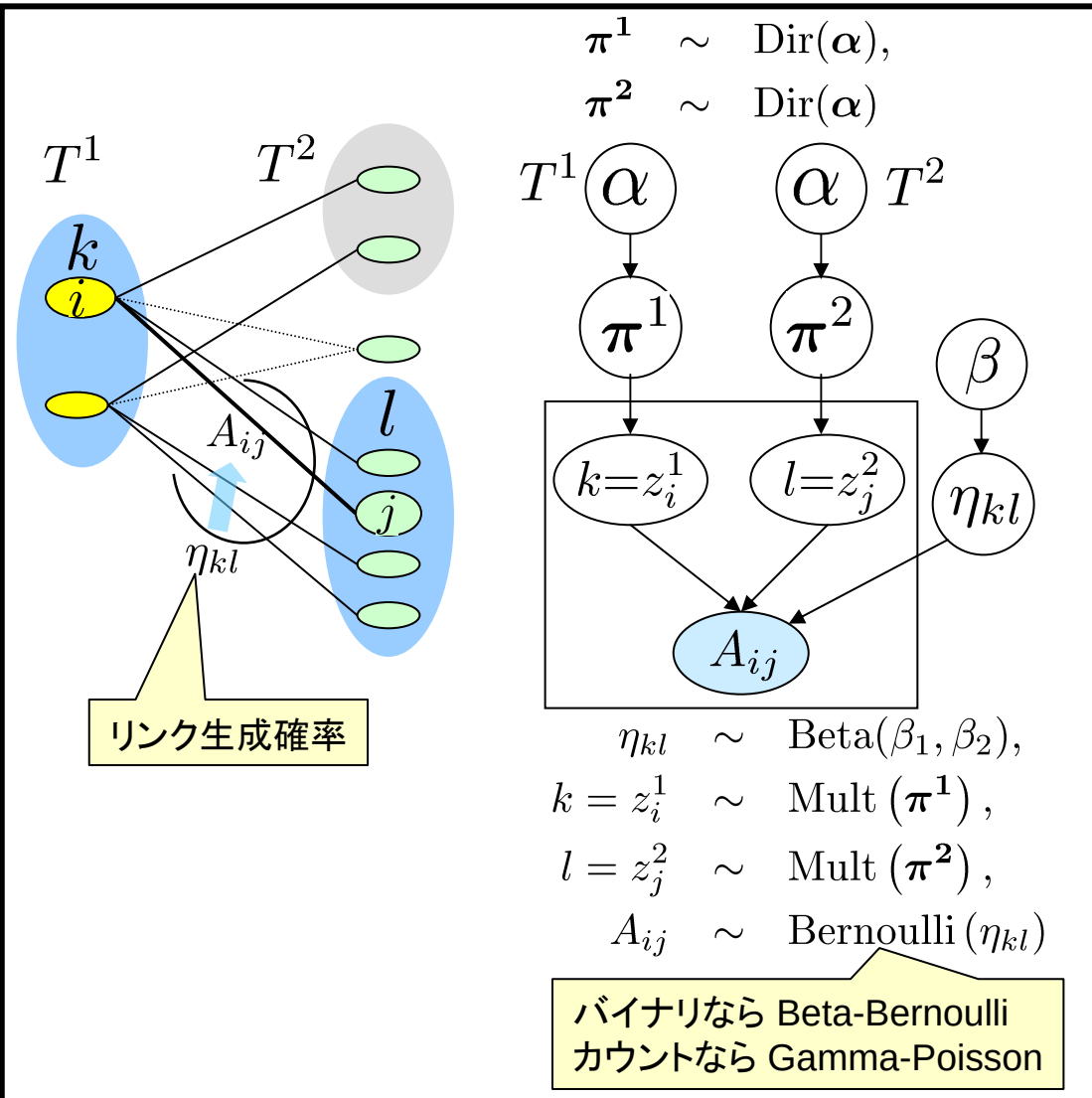


Mixture model



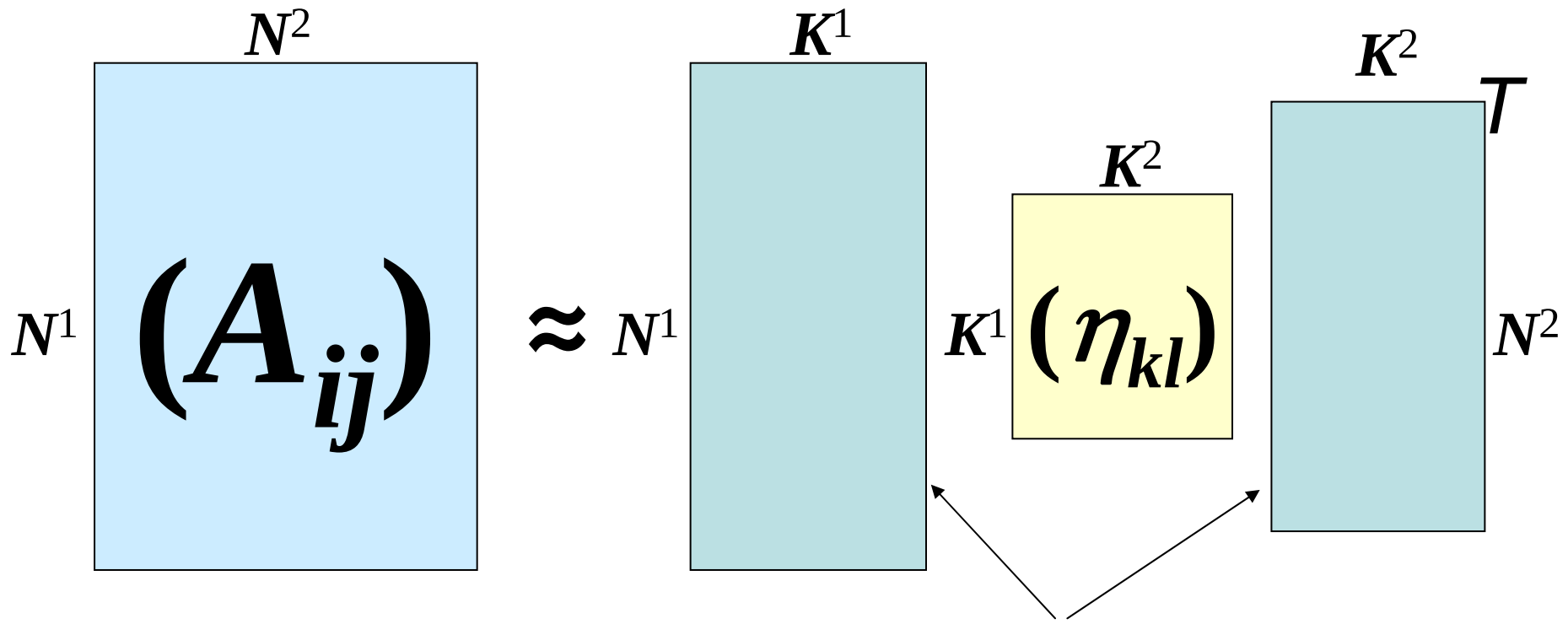
Infinite Mixture Model

同じクラスのノードは、みな
同じ相手にリンクを張る



ブロックモデルにおける次元圧縮

$N^1 \times N^2$ 次元の行列 A を $K^1 \times K^2$ 次元に「次元圧縮」しているとみなせる

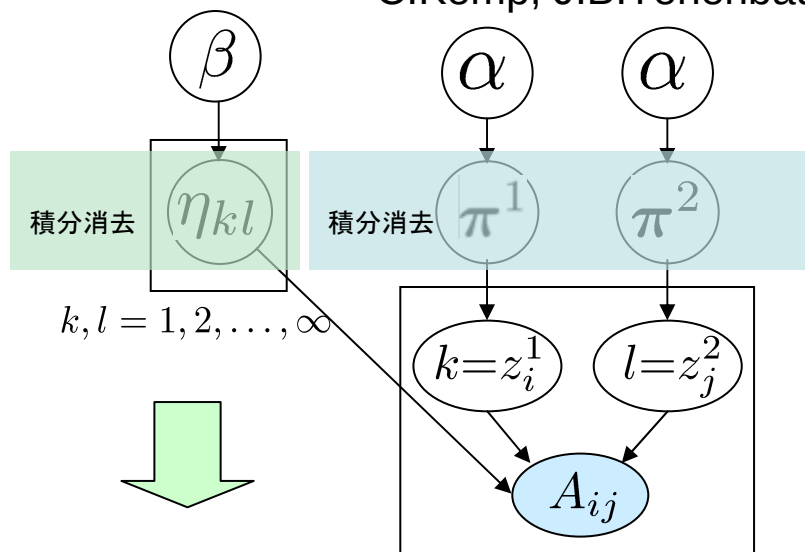


クラス所属確率 or どのクラスに所属するかを指定する行列

IRM (無限関係モデル)

ブロックモデルの「無限」化

Learning Systems of Concepts with an Infinite Relational Model,
C.Kemp, J.B.Tenenbaum, T.Yamada, N.Ueda, AAAI 2006 381-388



Chinese Restaurant Process

$$Z^1 | \alpha \sim \text{CRP}(\alpha)$$

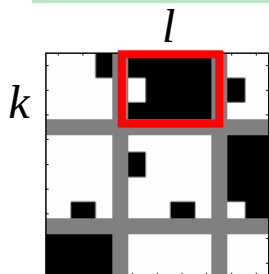
$$Z^2 | \alpha \sim \text{CRP}(\alpha)$$

$$\eta_{kl} | \beta \sim \text{Beta}(\beta_1, \beta_2)$$

$$A_{ij} | Z, \eta \sim \text{Bernoulli}(\eta_{kl})$$

$$P(A | Z^1, Z^2, \beta) = \int P(A | Z^1, Z^2, \eta) p(\eta | \beta) d\eta$$

$$= \prod_{k,l} \frac{B(m(k,l) + \beta_1, \bar{m}(k,l) + \beta_2)}{B(\beta_1, \beta_2)}$$



リンク数
□の黒の個数

非リンク数
□の白の個数

$$B(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

: ベータ関数

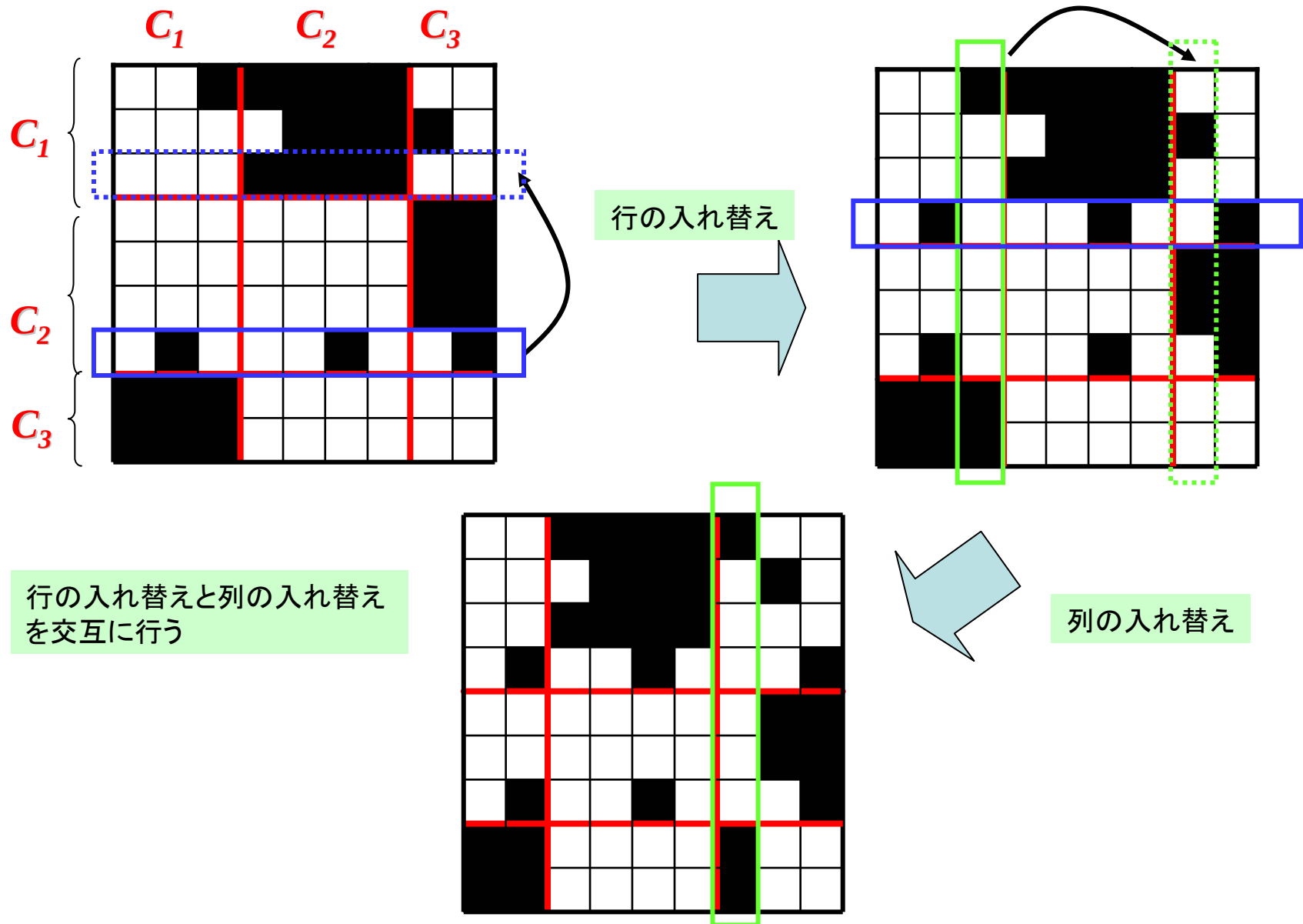
$$\text{Beta}(\theta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

: ベータ分布

joint

$$P(Z^1, Z^2 | A, \beta, \alpha) \propto \overbrace{P(Z^1 | \alpha) P(Z^2 | \alpha) P(A | Z^1, Z^2, \beta)}^{\text{joint}} \longrightarrow \max$$

Collapsed Gibbs Sampling

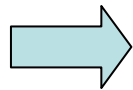


UMLS 生物医学オントロジデータのクラスタリング

UMLSデータ: 135の概念と49の述語から構成

概念(T1): Disease or Syndrome, Diagnostic Procedure, “Mammal”,

述語(T2): “complicates”, “affects”, “causes”, ...

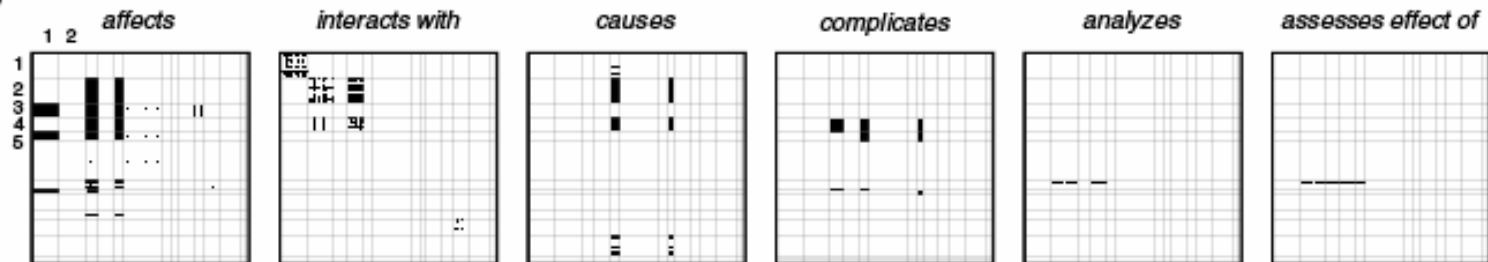


T1 x T1 x T2 のデータとして IRMを適用

a)

1.Organism	2.Chemicals	3.Biological functions	4.Bio-active substances	5.Diseases	affects	analyzes assesses effect of measures
Alga Amphibian Animal Archaeon Bacterium Bird	Amino Acid Carbohydrate Chemical Eicosanoid Isotope Steroid	Biological function Cell function Genetic function Mental process Molecular function Physiological function	Antibiotic Enzyme Poisonous substance Hormone Pharmacologic substance Vitamin	Cell dysfunction Disease Mental dysfunction Neoplastic process Pathologic function Expt. model of disease	diagnoses indicates prevents treats	carries out exhibits performs

b)



- biological functions **affect** organisms
- chemicals **cause** diseases
- biologically active substances **complicate** diseases

Erdős-Renyi mixture for graphs + DP

Jean-Jacques Daudin, Franck Picard, Stephane Robin, INRIA technical report

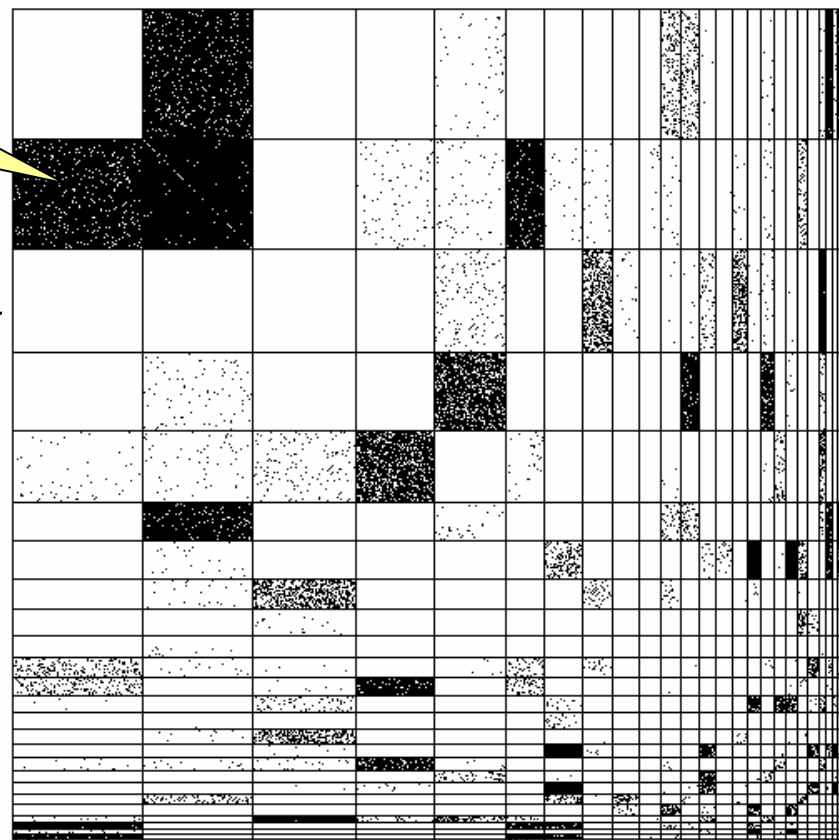
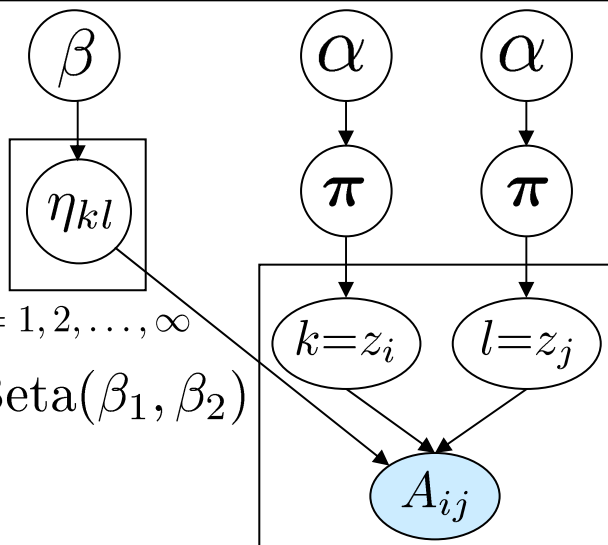
2次元のブロックモデルにおいて、 $T1=T2$ の時、ネットワーク
(グラフ)のノードクラスタリングと解釈できる

ブロック内、ブロック間では、Erdős-Renyi
random graph を仮定

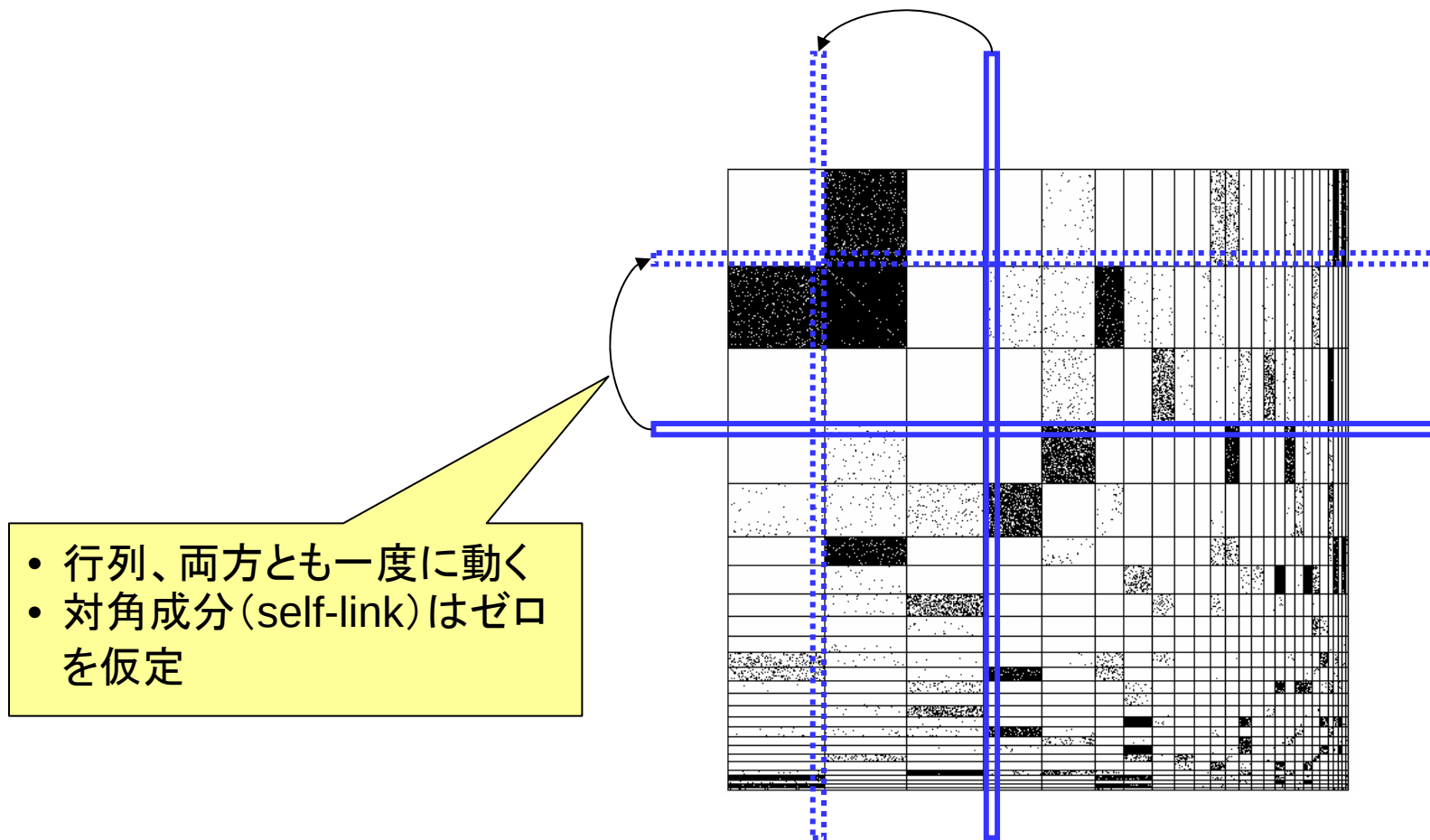
$$A_{ij}|Z, \eta \sim \text{Bernoulli}(\eta_{kl})$$

ブロックの分布は
CRP ($\approx$ Preferential Attachment)
を仮定

$$Z|\alpha \sim \text{CRP}(\alpha)$$



Collapsed Gibbs Sampling



A Bayesian Approach to Network Modularity + DP

Jake M. Hofman and Chris H. Wiggins, Phys. Rev. Lett. 100, 258701 (2008)

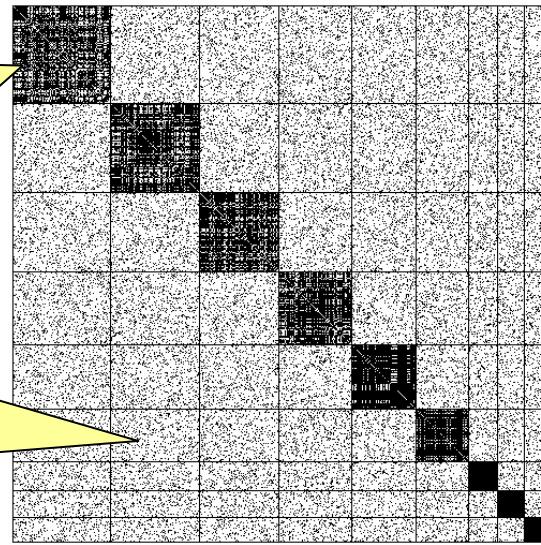
「対角成分」と「非対角成分」の二種類の Bernoulli パラメータ
(のみ)で、リンクが生成されると考える

対角成分のカウントはすべての
ブロックに関して和をとり、
まとめる

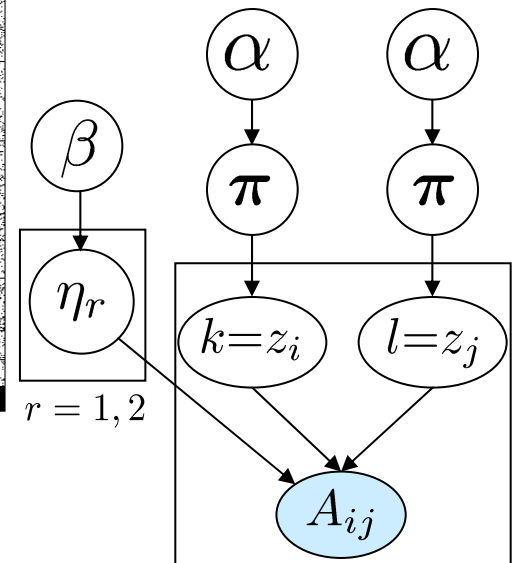
$$e = \sum_i e_{ii}$$

非対角成分のカウントについ
ても、すべてまとめる

$$a = \sum_{i \neq j} e_{ij}$$



$$Z|\alpha \sim \text{CRP}(\alpha)$$



$$\eta_1|\beta \sim \text{Beta}(\beta_1, \beta_2)$$

$$\eta_2|\beta \sim \text{Beta}(\beta_3, \beta_4)$$

$$A_{ij}|Z, \eta \sim \text{Bernoulli}(\eta_1) \text{ if } i = j$$

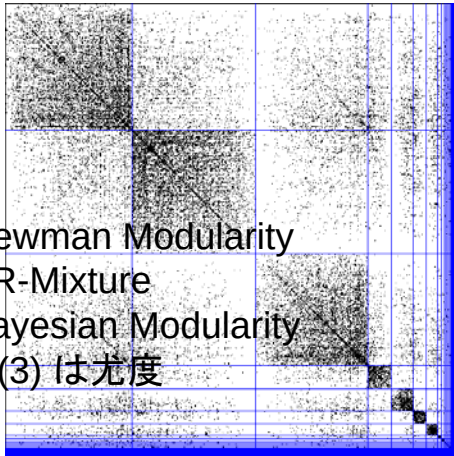
$$A_{ij}|Z, \eta \sim \text{Bernoulli}(\eta_2) \text{ if } i \neq j$$

$$P(A|Z, \beta) = \frac{B(e + \beta_1, \bar{e} + \beta_2)}{B(\beta_1, \beta_2)} \frac{B(a + \beta_3, \bar{a} + \beta_4)}{B(\beta_3, \beta_4)}$$

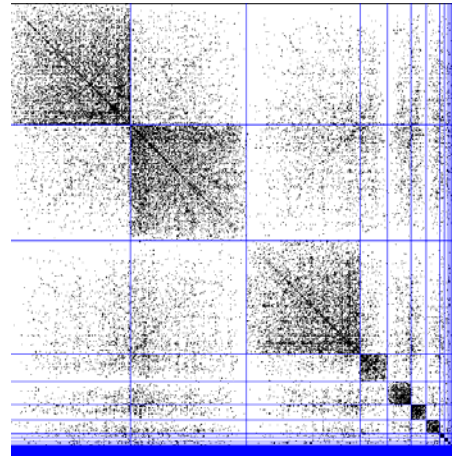
- 目的関数が modularity に近い
- resolution limit の問題を回避
- パラメータ数が少ないため、高速

単語連想ネットワークのクラスタリング例

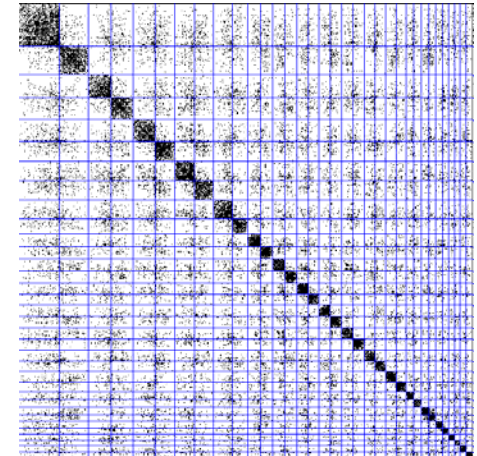
- (1) Newman Modularity
(2) ER-Mixture
(3) Bayesian Modularity
★ (2) (3) は尤度



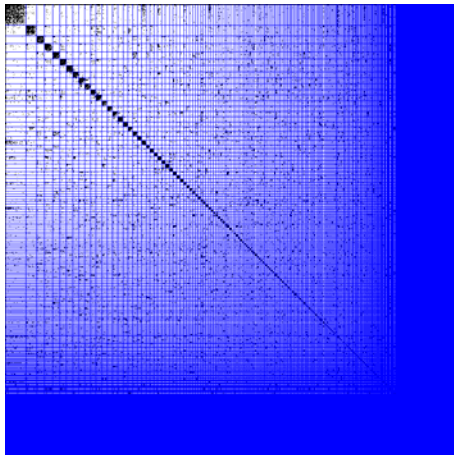
(1) 0.458846
(2) -229109.434651
(3) -227065.176490



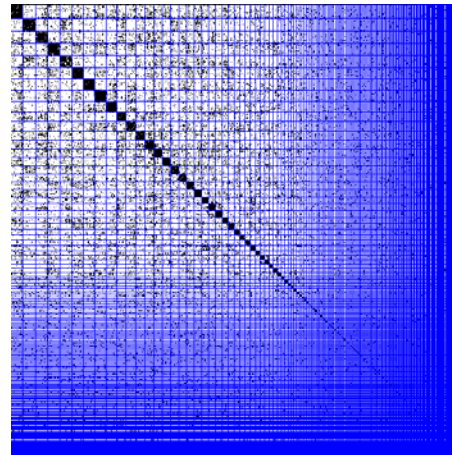
(1) 0.469602
(2) -229454.252000
(3) -226995.589016



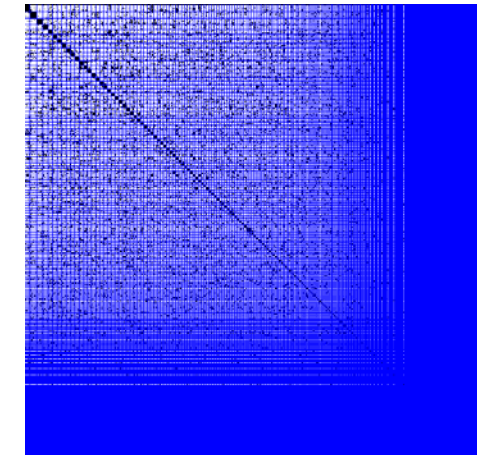
(1) 0.451034
(2) -216655.963400
(3) -215454.118461



(1) 0.366086
(2) -544871.856043
(3) -210315.793002

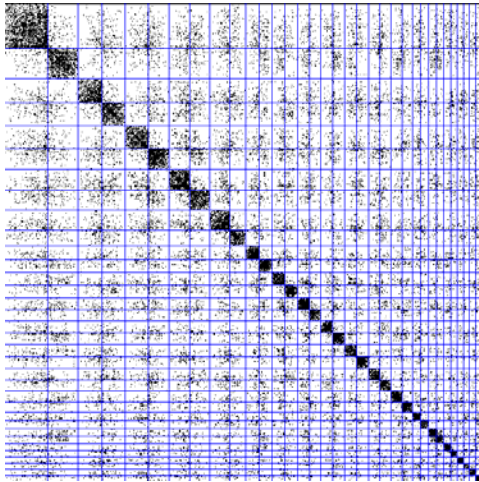


(1) 0.416832
(2) -302196.556078
(3) -205114.768975

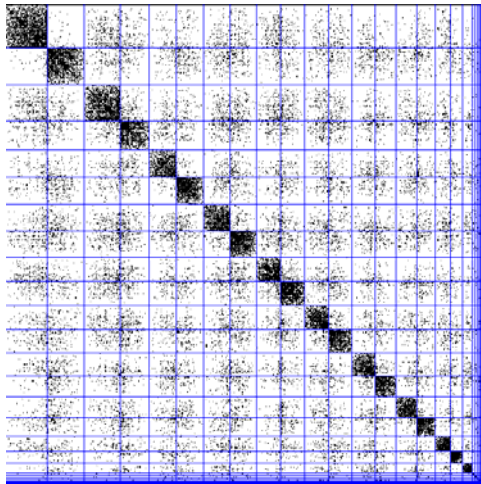


(1) 0.407359
(2) -462895.811110
(3) -196126.488926

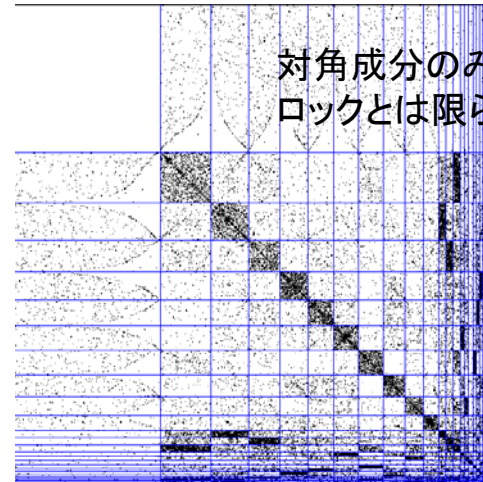
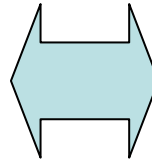
単語連想ネットワークのクラスタリング例(2)



(1) 0.451034
(2) -216655.963400
(3) -215454.118461

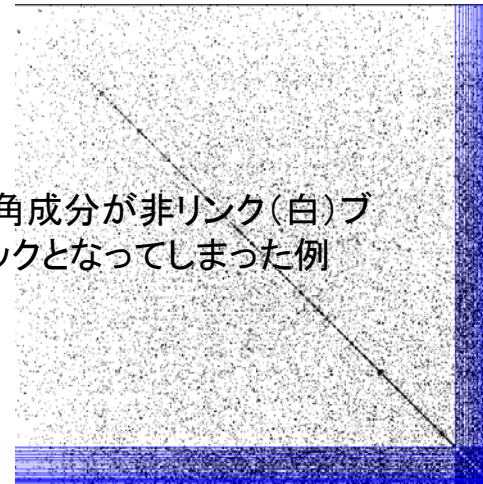


(1) 0.519364
(2) -214157.548246
(3) -234105.801714



対角成分のみリンク(黒)ブロックとは限らない

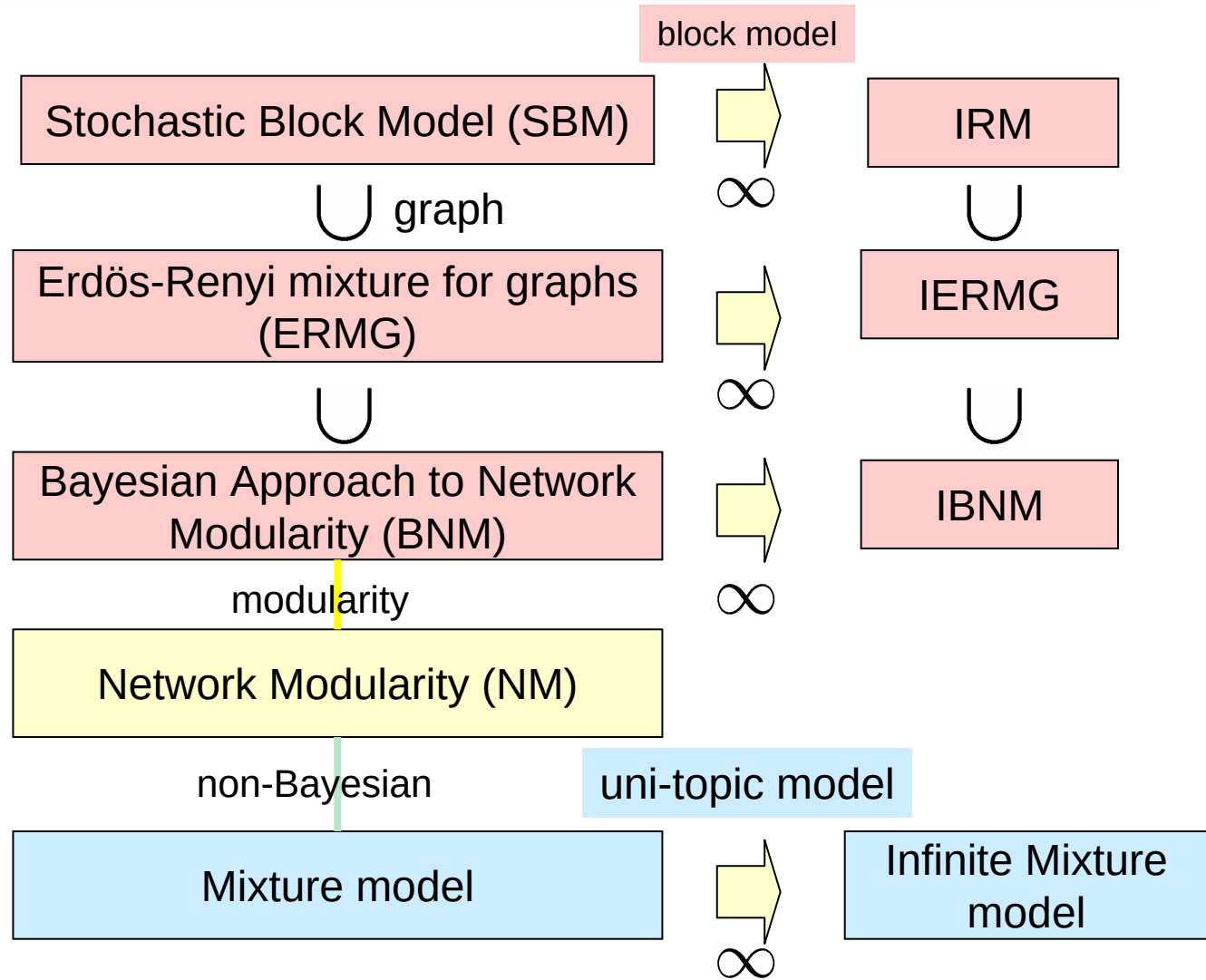
(1) 0.18102
(2) -196019.767261
(3) -243797.817000



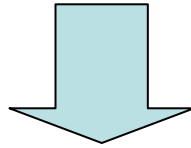
対角成分が非リンク(白)ブロックとなってしまう例

(1) -0.043643
(2) -231381.906832
(3) -230924.126523

ここまでのまとめ



ノードのクラスタリング

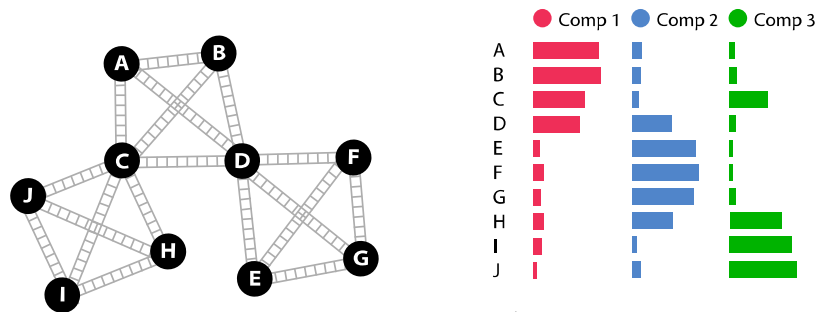


リンクのクラスタリング

ICMc: Interaction Component Model for Communities

Inferring vertex properties from topology in large networks,
J.Sinkkonen, J.Aukia and S. Kaski

無向リンクのクラスタリング (bag of links)



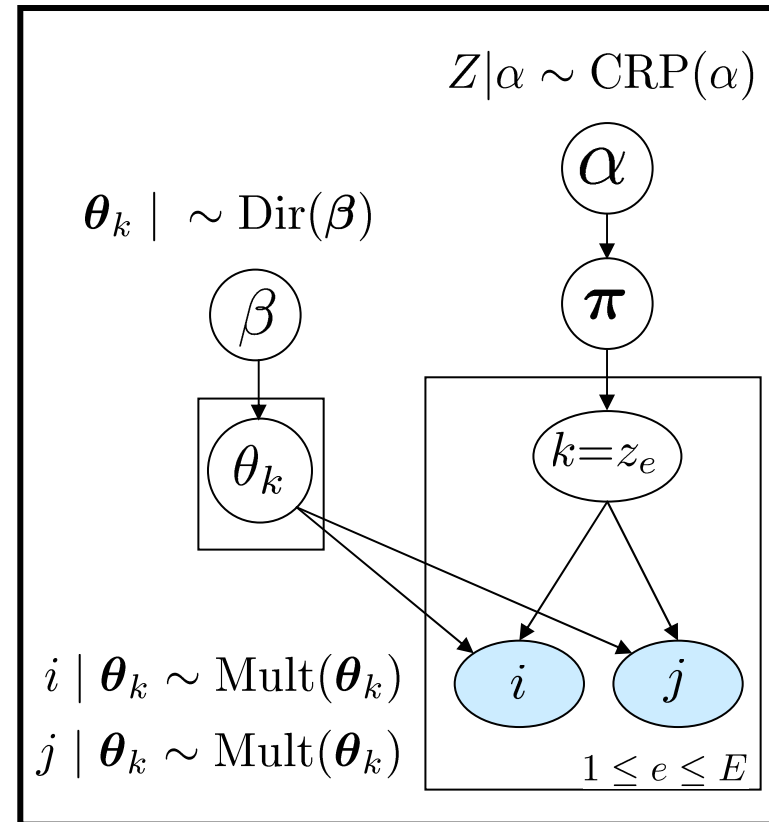
component wise
node degrees

link counts in z

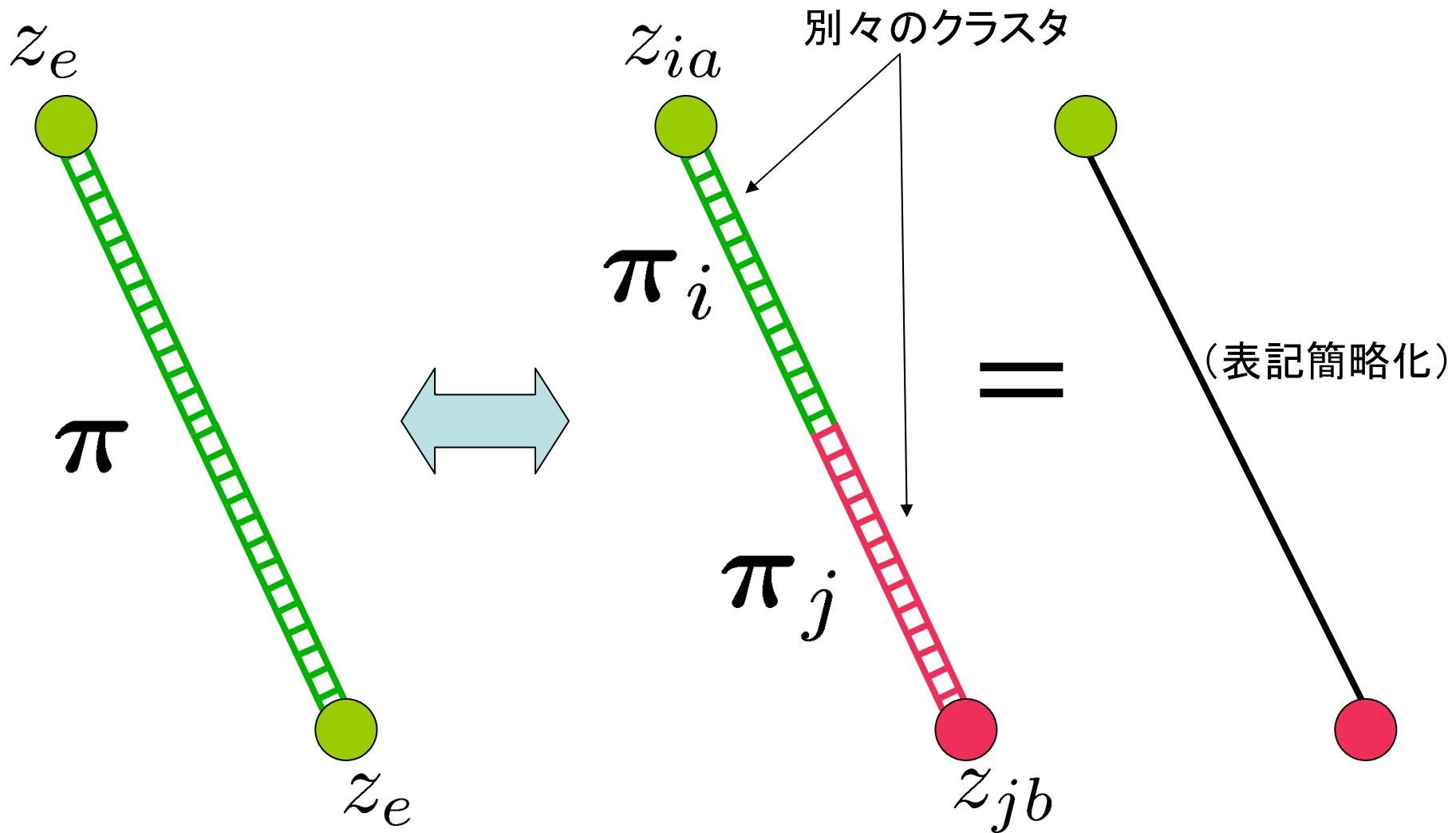
$$P(z|i, j) = \frac{n_{zi} + \beta}{2n_z + 1 + N\beta} \frac{n_{zj} + \beta}{2n_z + N\beta} \frac{n_z \text{ or } \alpha}{E - 1 + \alpha}$$

of nodes

of links



リンクのクラスタリングからリンクの両端のクラスタリングへ

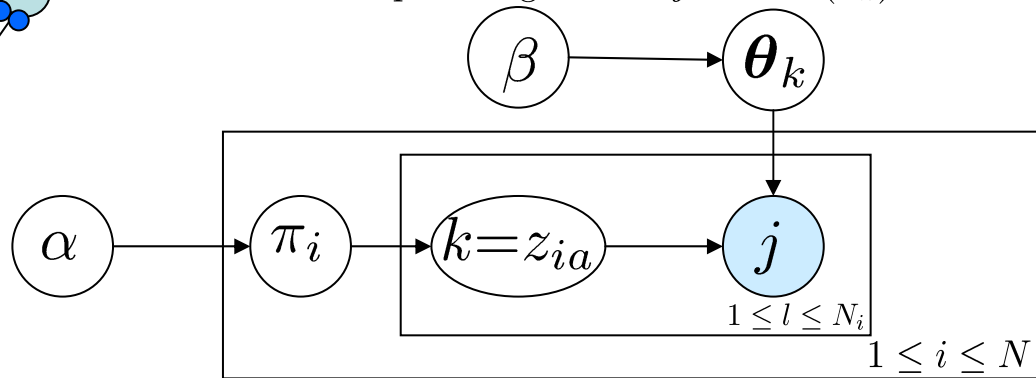
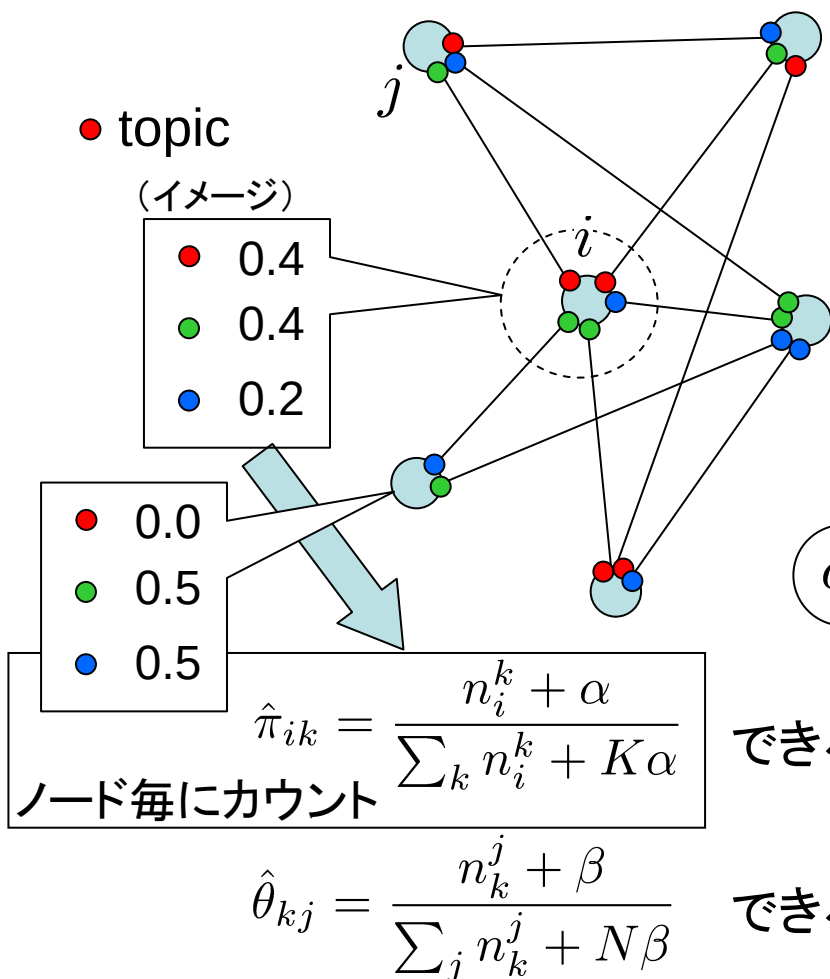


SSN-LDA: LDA-based community discovery approach

An LDA-based community structure discovery approach for large-scale social networks, H.Zhang, B.Qui, C.L.Giles, H.C.Foley and J.Yen

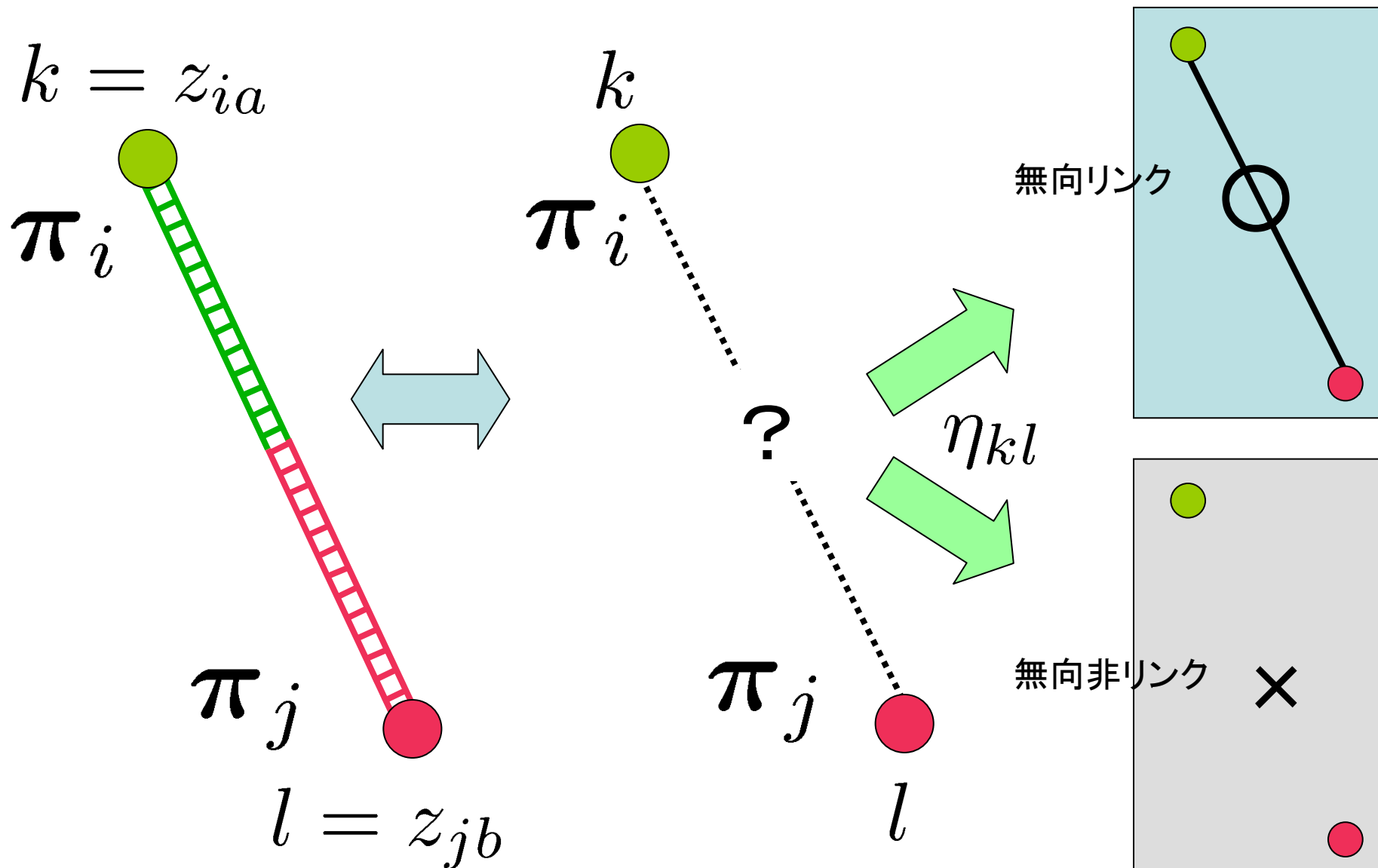
LDA を用いた、リンクのクラスタリング

1. Sample mixture components $\theta_k \sim \text{Dir}(\beta)$ for $k \in [1, K]$
2. For each node $i \in [1, N]$
 - (a) Sample $\pi_i \sim \text{Dir}(\alpha)$
 - (b) (Sample $N_i \sim \text{Poisson}(\xi)$)
 - (c) For each link a of the N_i links :
 - i. Sample a community $k=z_{ia} \sim \text{Mult}(\pi_i)$;
 - ii. Sample a target node $j \sim \text{Mult}(\theta_k)$



できるだけ少数のトピック/~~文書~~ノード
 ⇕
 できるだけ少数の~~単語~~トピックノード

リンクの両端のクラスタリングから全ノードペアのクラスタリングへ



Mixed membership stochastic block models

- Mixed membership stochastic block models for relational data with application to protein-protein interactions, E.M.Airoldi, D.M.Blei, S.E.Fienberg and E.P.Xing

1. For each node $i \in [1, N]$

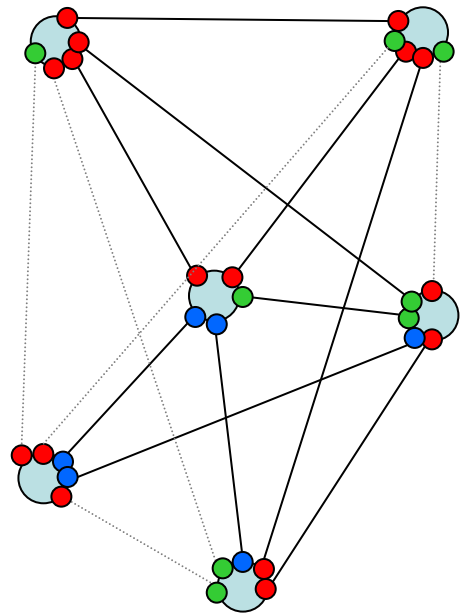
(a) Sample $\pi_i \sim \text{Dir}(\alpha)$ ← **LDA**

2. For each pair of nodes $(i, j) \in [1, N] \times [1, N]$

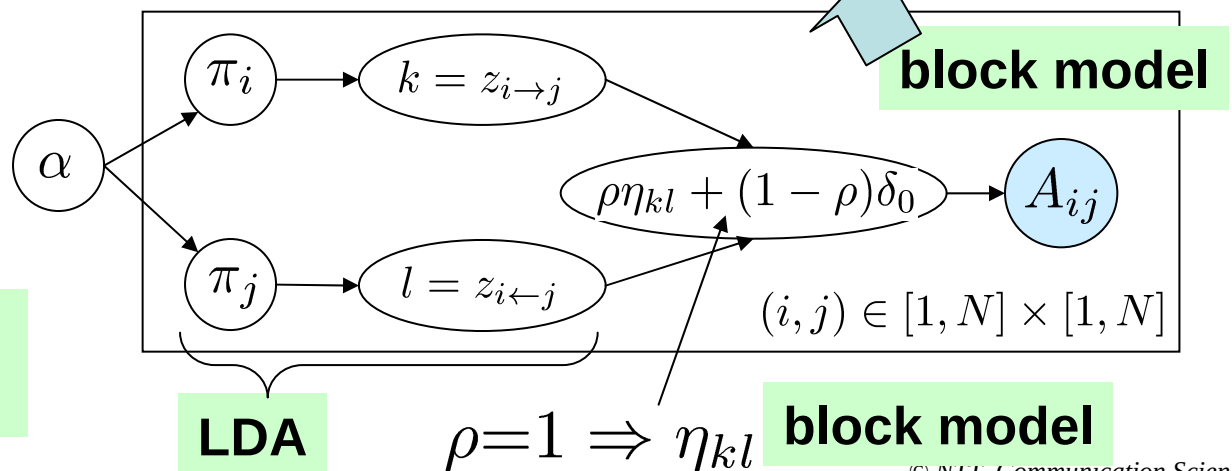
(a) Sample $k = z_{i \rightarrow j} \sim \text{Mult}(\pi_i)$ ← **LDA**

(b) Sample $l = z_{i \leftarrow j} \sim \text{Mult}(\pi_j)$

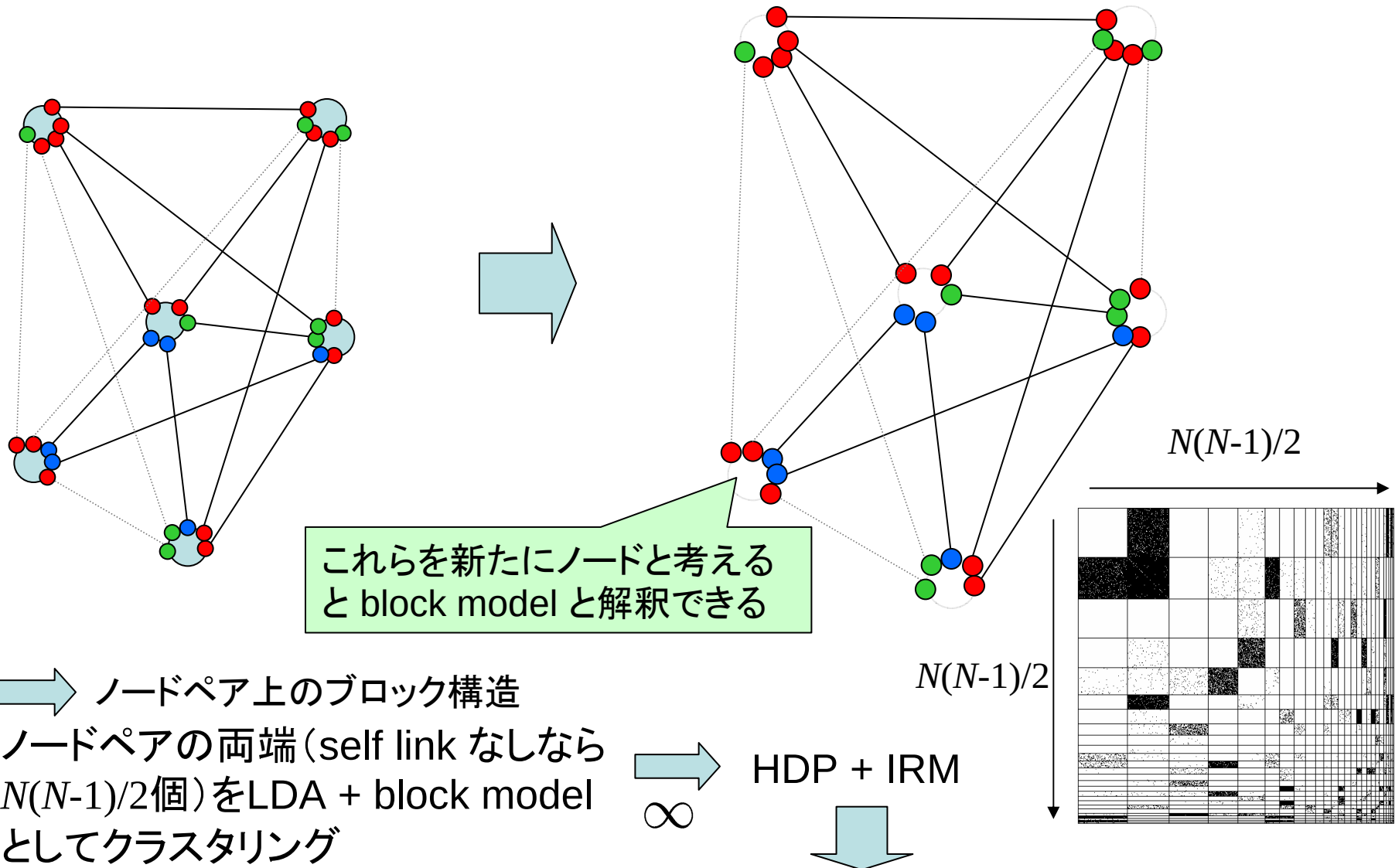
(c) Sample $A_{ij} \sim \text{Bernoulli}(\rho\eta_{kl} + (1 - \rho)\delta_0)$



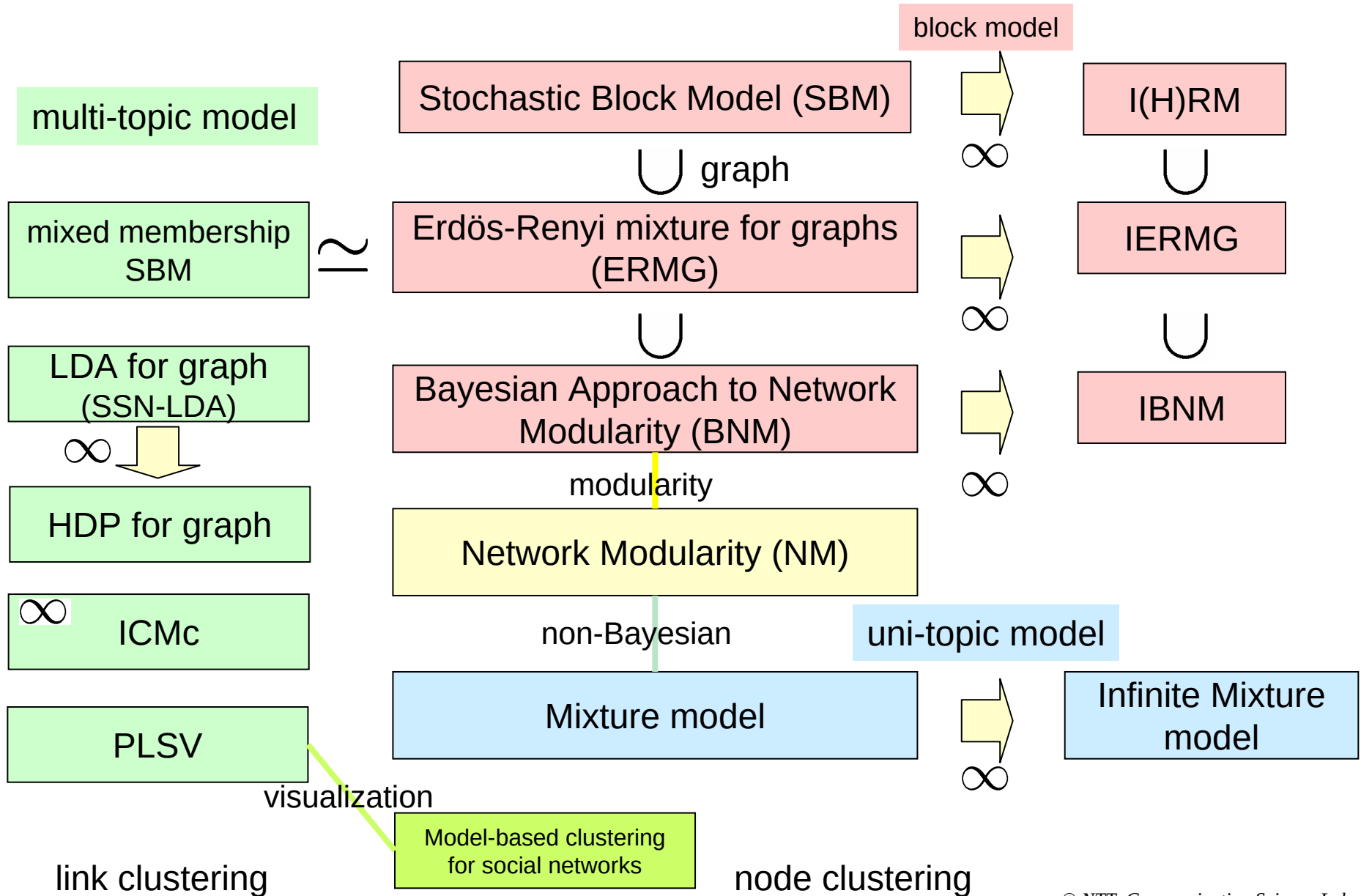
**LDA と block model
の融合**



Mixed membership stochastic block models (2)



まとめ



主な文献リスト

- Learning Systems of Concepts with an Infinite Relational Model, C.Kemp, J.B.Tenenbaum, T.Yamada, N.Ueda, AAAI 2006 381-388
- A mixture model for random graphs, Jean-Jacques Daudin, Franck Picard, Stephane Robin, INRIA technical report
- A Bayesian Approach to Network Modularity, Jake M. Hofman and Chris H. Wiggins, Phys. Rev. Lett. 100, 258701 (2008)
- Learning infinite hidden relational models, Z. Xu, V. Tresp, K. Yu and H.P. Kriegel: 本質的に、IRM と同じ
- Inferring vertex properties from topology in large networks, J.Sinkkonen, J.Aukia and S. Kaski: もっとも単純なリンククラスタリング
- An LDA-based community structure discovery approach for large-scale social networks, H.Zhang, B.Qui, C.L.Giles, H.C.Foley and J.Yen: LDA を用いた、リンクのクラスタリング
- Knowledge discovery of semantic relationships between words using nonparametric Bayesian graph model, I.Sato, M.Yoshida and H.Nakagawa: LDA を「無限化」してHDPを用いた
- Mixed membership stochastic block models for relational data with application to protein-protein interactions, E.M.Airoldi, D.M.Blei, S.E.Fienberg and E.P.Xing: 一種のブロックモデル
- Model-based clustering for social networks, M.S.Handcock and A.E.Raftery: 「可視化」を意識
- Probabilistic Latent Semantic Visualization Topic Model for Visualizing Documents, T.Iwata, T.Yamada and N.Ueda: 「可視化」を意識、PLSIに近い