Minimum Dominating Set Approach to Analysis and Control of Biological Networks

Tatsuya Akutsu

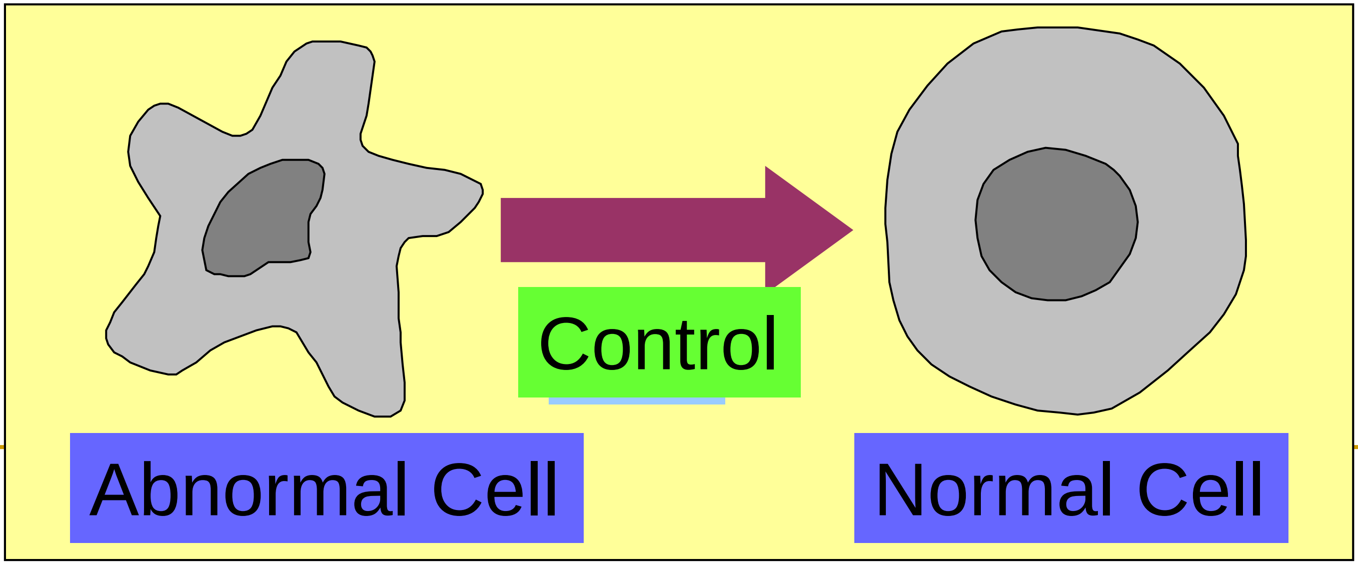
Bioinformatics Center

Institute for Chemical Research, Kyoto University

Joint work with Jose Nacher in Toho University

Motivation: Control Theory for Biological Systems

- One of the main targets of Systems Biology
  - Though control theory is well established for linear systems, biological systems have non-linear components and are very complex (large-scale)
  - May lead to new drugs and treatment methods
- Practical control methods exist, but no useful theory
  - Introduction of 4 genes turns normal cells into induced pluripotent stem cells (iPS cells)



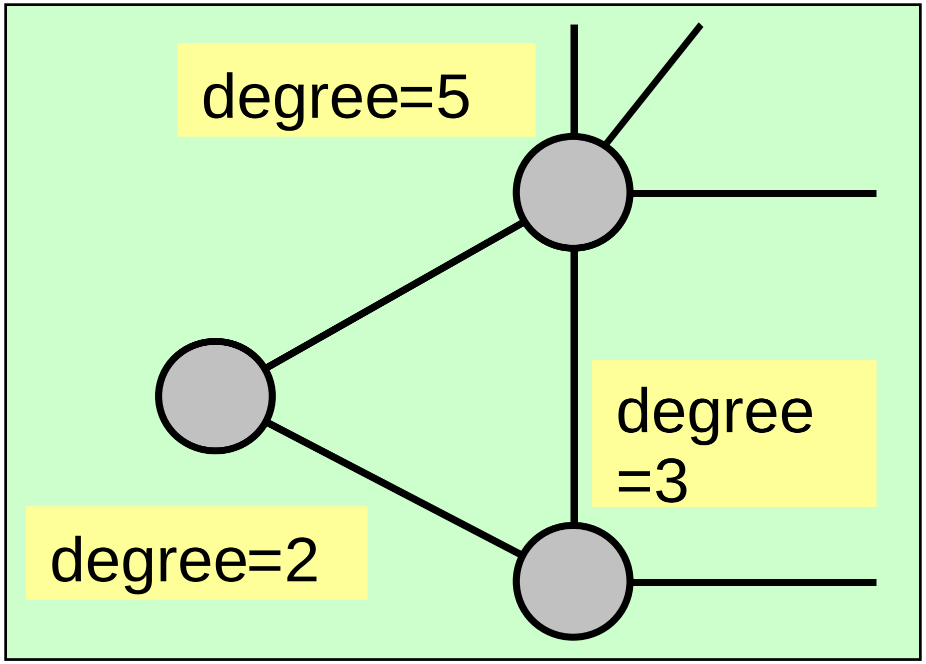
Contents

- Scale-free Networks
- Controllability in Scale-free Networks
- Minimum Dominating Set (MDS)
  - Relation to Structural Controllability
  - Theoretical Analysis of MDS Size
  - Computer Simulation
  - Database Analysis
- Applications to Analysis of Biological Networks
- Extensions
- Conclusion

Scale-free Networks

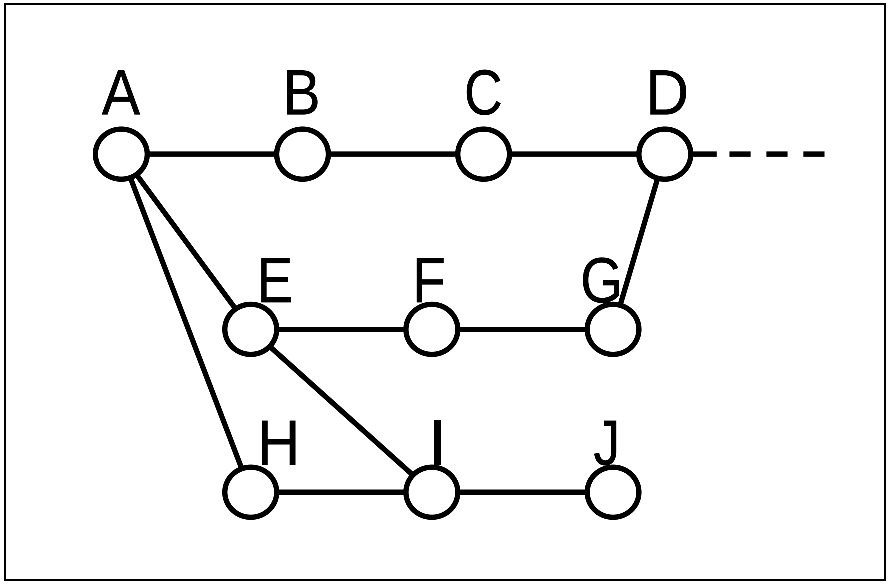
Scale-Free Network [Barabasi & Albert, 1999]

- Degree of a node
  - The number of adjacent nodes
- $P(k)$
  - Degree distribution
  - Frequency of nodes with degree k
- Scale-free network
  - $P(k)$ follows power law
  - Different from random networks



$$P(k) \propto k^{-\gamma}$$

Metabolic Network, Graph and Degree



- Degree
  - Node with degree 1: J
  - Nodes with degree 2: B, C, D, F, G, H
  - Nodes with degree 3: A,E, I
- $P(k)$ (degree distribution):

  $P(1)=0.1$, $P(2)=0.6$, $P(3)=0.3$, $P(4)=P(5)=P(6)=\ldots=0$

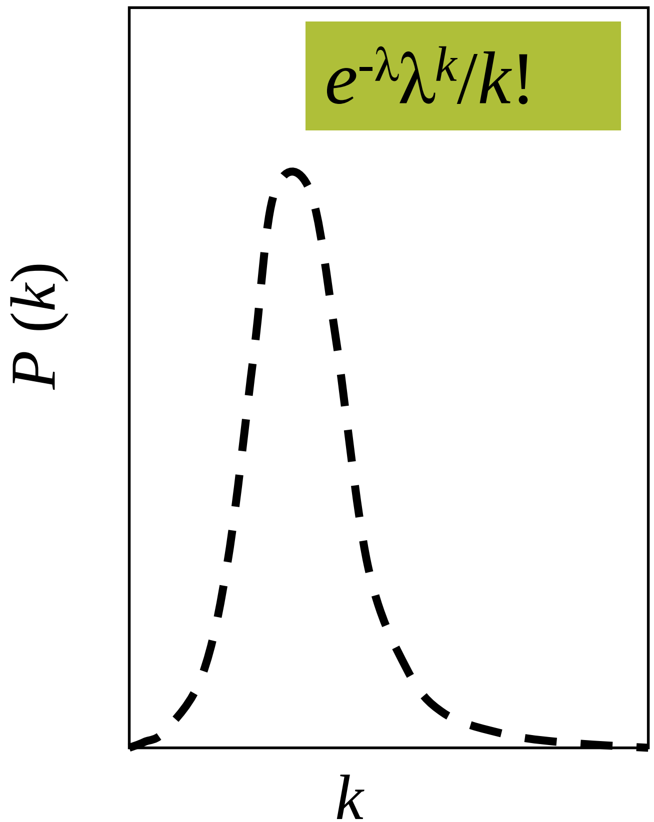
Scale-Free Distribution

$$P(k) \propto k^{-\gamma}$$

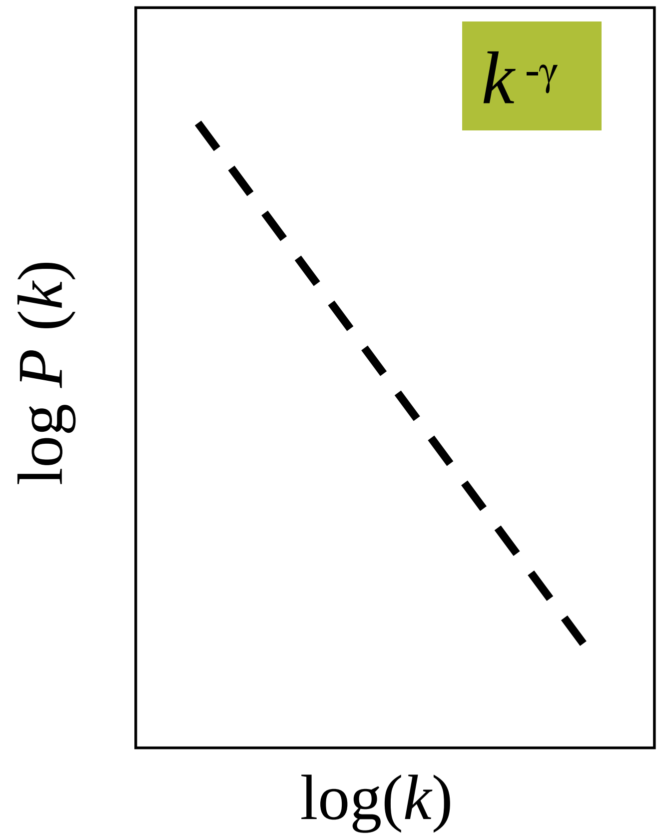
- Power laws are scale free because if k is rescaled (multiplied by a constant), then $P(k)$ is still proportional to $k^{-\gamma}$
- Many real networks (e.g., genetic networks, metabolic networks, protein-protein interaction networks) are reported to have the scale-free property

Poisson Distribution and Power-Law Distribution

Poisson distribution (random graph)



Power-law distribution (scale-free graph)



Controllability in Scale-free Networks

[Liu, Slotine, Barabasi: Nature, 2011]

Controllability of Linear Systems (1)

Input:

- Linear System: $$\frac{d\mathbf{x}(t)}{dt} = A\mathbf{x}(t) + B\mathbf{u}(t)$$
- Initial state: $\mathbf{x}_0$ Final state: $\mathbf{x}_F$

Output:

- $\mathbf{u}(t)$ (function of t) which drives the system from $\mathbf{x}_0$ to $\mathbf{x}_F$ in finite time

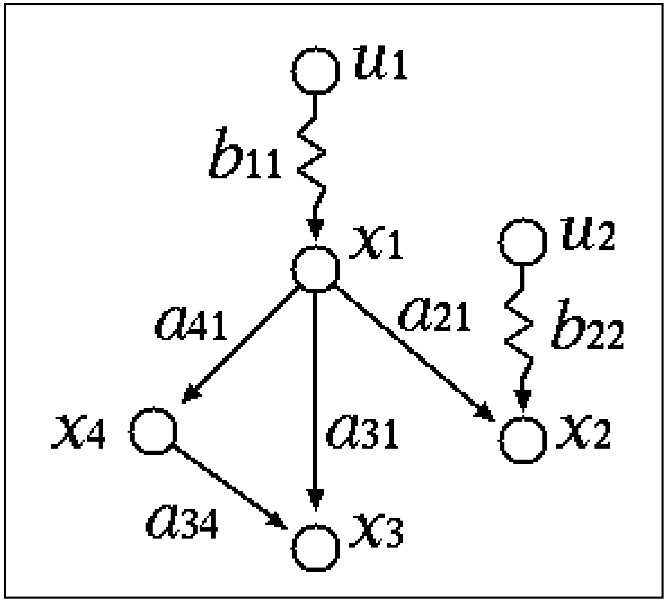
$\mathbf{x}(t)$: N-dim. real vector (internal nodes)
$\mathbf{u}(t)$: M-dim. real vector (control nodes)
A: $N \times N$ real matrix
B: $N \times M$ real matrx

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_N \end{pmatrix} = A \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix} + B \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_M \end{pmatrix}$$

Controllability of Linear Systems (2)

Fact. System is controllable iff

$N \times NM$ matrix $C=(B,AB,A^2B,\ldots,A^{N-1}B)$ has full rank (i.e., $\mathrm{rank}(C)=N$).



$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ a_{21} & 0 & 0 & 0 \\ a_{31} & 0 & 0 & a_{34} \\ a_{41} & 0 & 0 & 0 \end{pmatrix} \qquad B = \begin{pmatrix} b_{11} & 0 \\ 0 & b_{22} \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

$$C = \begin{pmatrix} b_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & b_{22} & a_{21}b_{11} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_{31}b_{11} & 0 & a_{34}a_{41}b_{11} & 0 & 0 & 0 \\ 0 & 0 & a_{41}b_{11} & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$\mathrm{rank}(C)=4$ for most parameters ⇒ **structural controllability**

Structural Controllability

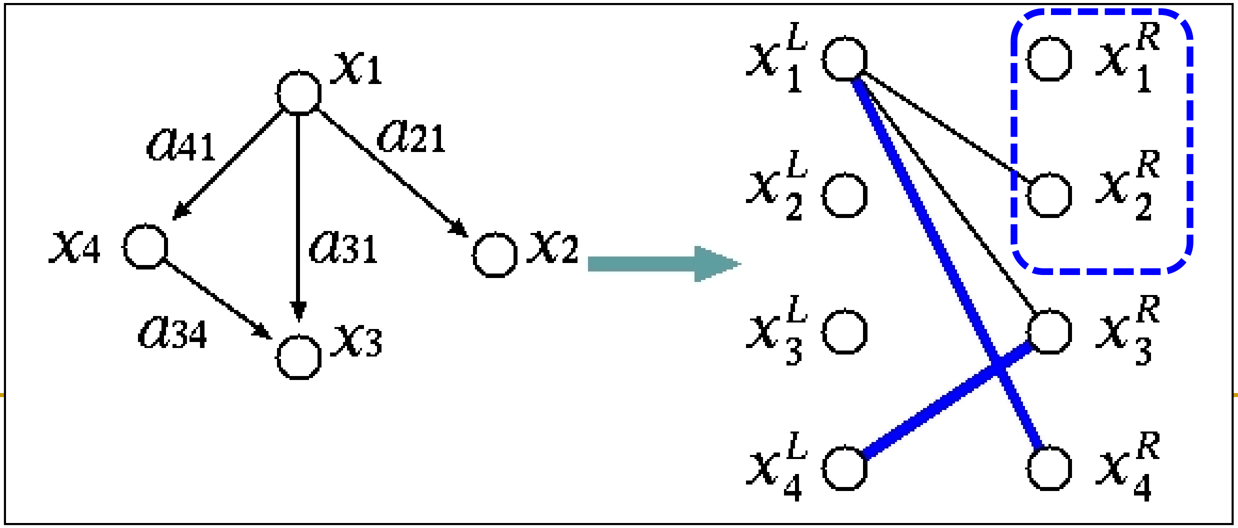
$G_B(V^L, V^R; E_B)$: bipartite graph constructed from $G(V,E)$ by

$$V^L = \{x_i^L \mid x_i \in V\},\ V^R = \{x_i^R \mid x_i \in V\},\ (x_i, x_j) \in E \Leftrightarrow (x_i^L, x_j^R) \in E_B$$

Thm. [Liu et al. 2011]

The minimum number of nodes needed to fully control the system is max $\{N\text{-}M^*, 1\}$,

where M^* is the size of the maximum matching of G_B.



Controllability of Scale-free Networks

The number of needed driver nodes [Liu et al. 2011]

- Random networks:

$$N_D \approx n \cdot e^{-<k>/2}$$

- Scale-free networks

$$N_D \approx n \cdot \exp\left[-\tfrac{1}{2}\left(1-\tfrac{1}{\gamma-1}\right)< k >\right]$$

⇒ if $\gamma<2$, many nodes must be controlled

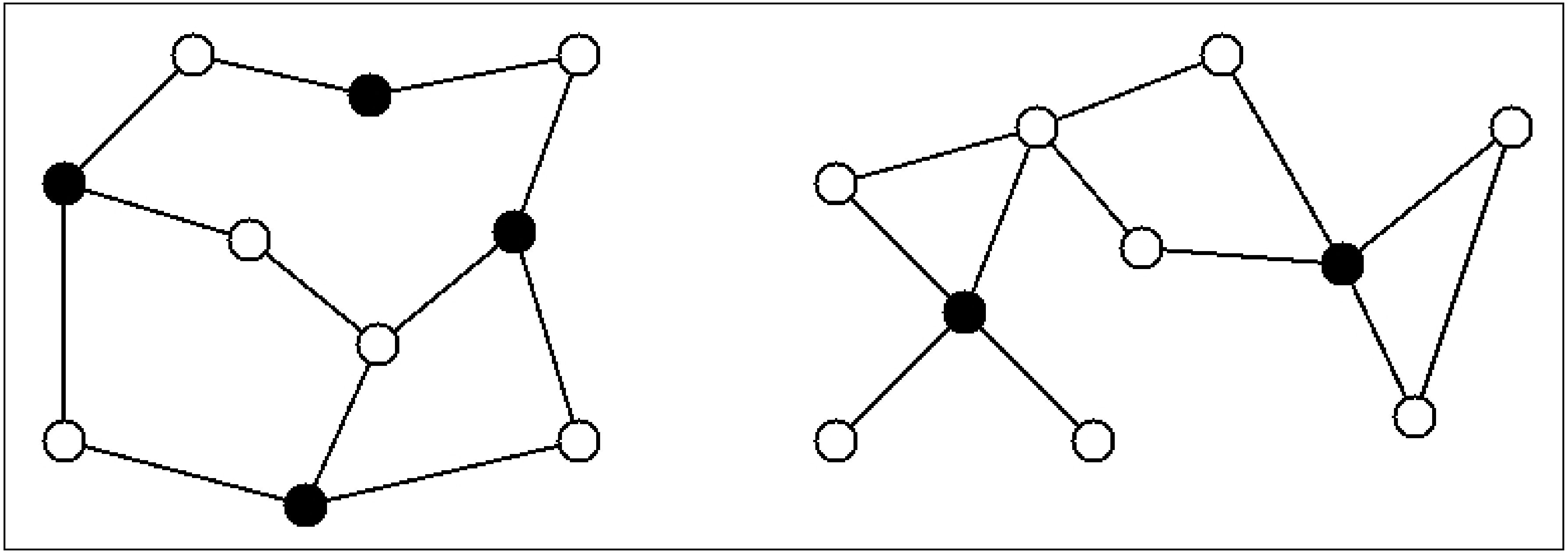
$<k>$: average degree

n: number of nodes in a network

Minimum Dominating Set and Its Relation to Structural Controllability

Minimum Dominating Set (1)

- V_D is a dominating set of undirected graph $G(V,E) \Leftrightarrow (\forall v \in V\text{-}V_D)(\exists u \in V_D)(\{u,v\} \in E)$
- Minimum dominating set: dominating set with the smallest number of nodes

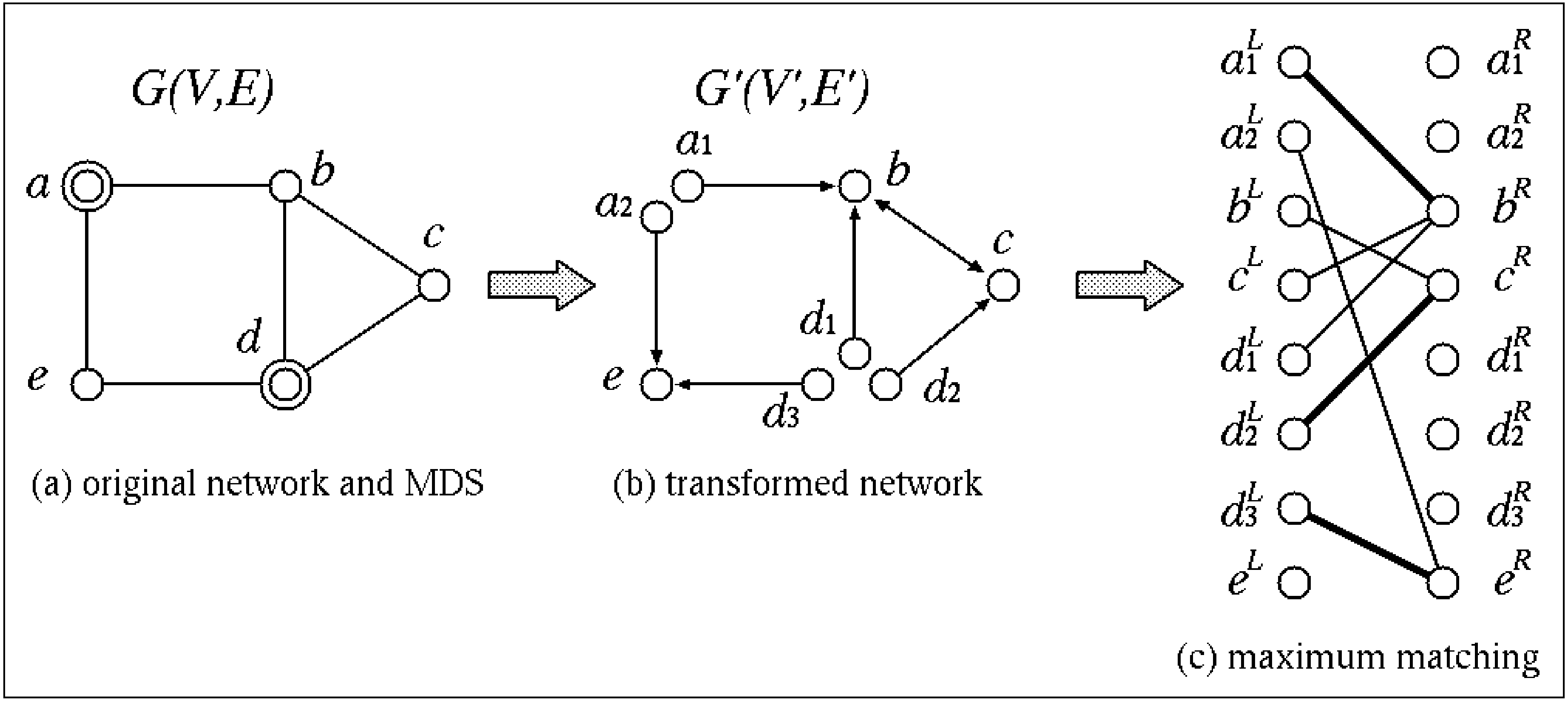


Minimum Dominating Set (2)

- Well-known concept in graph theory and computer science
- NP-hard, but can be solved exactly by using Integer Linear Programming (ILP) to some extent
- Has been applied to design/control of
  - mobile ad-hoc networks (MANET)
  - transportation routing
  - computer communication networks

Relation between MDS and Controllability

Thm. Suppose that every edge in a network is bi-directional and every node in MDS can control all of its outgoing links separately. Then, the network is structurally controllable by selecting the nodes in MDS as the driver nodes.



[Nacher & Akutsu: New J. Phys. 2012]

ILP-based Method for MDS

- Very simple, but works for networks with a few thousands of nodes in many cases

$$\min \sum_{i=1}^{n} x_i$$

$$s.t. \quad \sum_{\{j \mid j=i \vee \{v_i, v_j\} \in E\}} x_j \geq 1, \; i = 1, \ldots, n$$

$$x_i \in \{0,1\}$$

- $x_i = 1 \Leftrightarrow x_i$ in MDS

Theoretical Analysis of MDS Size

Estimation of MDS Size in Scale-free Networks

$\gamma>2$

- Upper bound: trivially $O(n)$
- Lower bound: $\Omega(n)$

$\gamma<2$

- Upper bound: $O(n^{1-(2-\gamma)(\gamma-1)})$
  - taking the minimum order $O(n^{0.75})$ when $\gamma=1.5$

Based on a kind of mean-field approximation

[Nacher & Akutsu: J. Phys.: Conf. Ser. 2013]

Lower Bound for γ>2 (1)

- Assuming $\alpha k^{-\gamma}$, we have

$$\alpha n \int_1^n k^{-\gamma} dk = \frac{\alpha n}{\gamma - 1}(1 - n^{-\gamma+1}) = n \;\Rightarrow\; \alpha \approx \gamma - 1$$

- The following is well known,

where $C(S)$ is the set of edges between S and V-S

if $|S|+|C(S)|<n$, S is not a dominating set

- If we select all nodes with degree > K, we have

$$|C(S)| < \alpha n \int_K^n k \cdot k^{-\gamma} dk \approx n(\gamma - 1)\int_K^n k^{-\gamma+1} dk$$

$$= n \cdot \left(\frac{\gamma - 1}{\gamma - 2}\right) \cdot \left(\frac{1}{K^{\gamma-2}} - \frac{1}{n^{\gamma-2}}\right) < n \cdot \left(\frac{\gamma - 1}{\gamma - 2}\right) \cdot \frac{1}{K^{\gamma-2}}$$

Lower Bound for γ>2 (2)

- Since we can assume $|S|<n/2$, we should have

$$n \cdot \left(\frac{\gamma - 1}{\gamma - 2} \right) \cdot \frac{1}{K^{\gamma - 2}} > n/2$$

- Then, we estimate a lower bound of $|S|$ by

$$|S| \approx \alpha n \int_K^n k^{-\gamma} dk \approx n \left(\frac{1}{K^{\gamma-1}} - \frac{1}{n^{\gamma-1}} \right)$$

$$\approx n \cdot \left(\frac{1}{K^{\gamma-1}} \right) > \left[2 \cdot \left(\frac{\gamma - 1}{\gamma - 2} \right) \right]^{-\frac{\gamma-1}{\gamma-2}} \cdot n$$

- This means that the number increases as γ increases

Upper Bound for γ<2 (1)

- We select all nodes with degree greater than $K=n^{\beta}$ as DS
- Then, N_{DS}=#nodes in DS (dominating set) is given by

$$N_{DS} = \alpha n \int_{n^{\beta}}^{n} k^{-\gamma} dk = n(n^{-\beta(\gamma-1)} - n^{-(\gamma-1)}) = O(n^{1-\beta(\gamma-1)})$$

- On the other hand, the total number of edges E_G is

$$E_G = \alpha n \int_{1}^{n} k \cdot k^{-\gamma} dk = \tfrac{\gamma-1}{2-\gamma} \cdot n \cdot (n^{2-\gamma} - 1)$$

- E_{DS} (=the number of edges covered by DS) is

$$E_{DS} = \alpha n \int_{n^{\beta}}^{n} k \cdot k^{-\gamma} dk = \tfrac{\gamma-1}{2-\gamma} \cdot n \cdot (n^{2-\gamma} - n^{\beta(2-\gamma)})$$

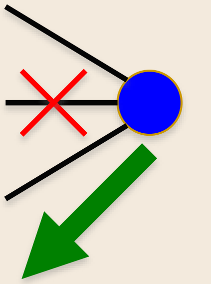
- Then, prob. that an arbitrary edge is NOT covered by DS is

$$\frac{E_G - E_{DS}}{E_G} = \frac{n^{\beta(2-\gamma)} - 1}{n^{2-\gamma} - 1} \approx n^{(\beta-1)(2-\gamma)}$$

Upper Bound for γ<2 (2)

- Since a node is covered by DS if at least one edge connecting to the node is covered by DS, the expected number ($N_{G\text{-}DS}$) of nodes not covered by DS is

$$N_{G-DS} \leq O(n \cdot n^{(\beta-1)(2-\gamma)}) = O(n^{1+(\beta-1)(2-\gamma)})$$



$\circ \in G-DS$

- Here, we balance $N_{G\text{-}DS}$ with N_{DS} by letting

$$1-\beta(\gamma-1) = 1+(\beta-1)(2-\gamma)$$

which results in $\beta=2-\gamma$.

- Therefore, an upper bound of the size of DS is estimated as

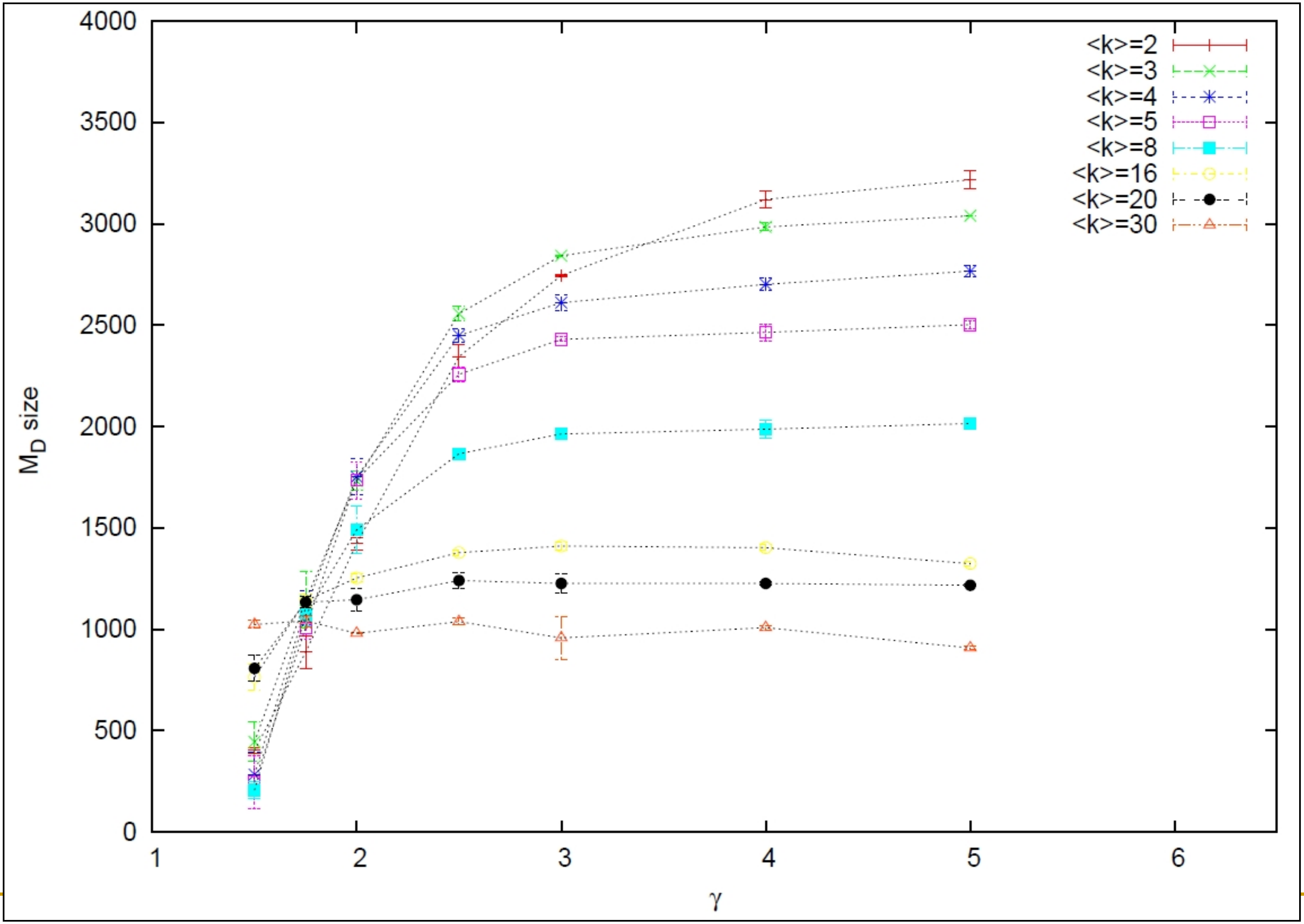
$$O(n^{1-(2-\gamma)(\gamma-1)})$$

which is $o(n)$ for $1<\gamma<2$

- It is interesting that it takes the minimum ($O(n^{0.75})$) when $\gamma=1.5$

Computer Simulation

MDS size vs. Scaling Exponent (γ)



- MDS size decays as γ decays (especially around γ=2)

Database Analysis

Data

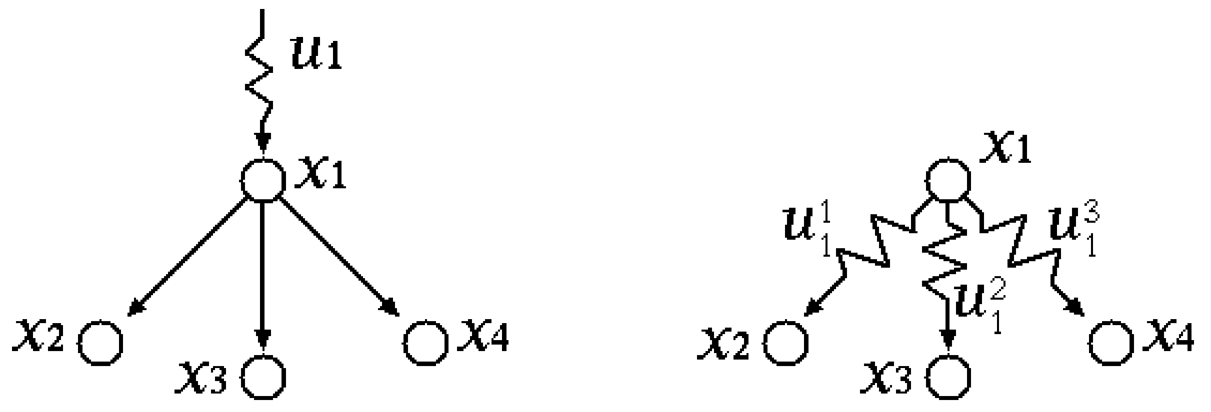
Name	Nodes	GCC	m_D	$< k >$	l	d	C	NC
PPI *C. elegans*	2,651	2,386	0.182	3.20	4.80	14	0.022	0.077
PPI *D. melanogaster*	7,498	7,351	0.199	6.14	4.40	12	0.012	0.023
PPI *E. coli*	1,865	1,447	0.229	8.12	3.81	12	0.109	0.109
PPI *H. sapiens*	1,607	805	0.239	2.92	6.53	19	0.107	0.042
PPI *M. musculus*	599	50	0.220	2.20	4.42	9	0.060	0.208
PPI *S. cerevisiae*	4,963	4,902	0.179	7.03	4.14	11	0.097	0.056
TRN *S. cerevisiae*	688	662	0.126	3.20	5.20	15	0.049	0.103
TRN *E. coli*	418	328	0.176	2.78	4.83	13	0.110	0.213
U.S. Airports	500	500	0.102	11.92	2.99	7	0.617	0.268
Word adjacency (*Japanese*)	2,704	2,698	0.109	5.92	3.07	8	0.220	0.267
Word adjacency (*Spanish*)	12,642	11,558	0.067	7.44	2.91	10	0.376	0.258
Collaboration (*ca-HepTh*)	9,877	8,638	0.205	5.74	5.94	18	0.482	0.007
Collaboration (*ca-GrQc*)	5,242	4,158	0.186	6.45	6.04	17	0.557	0.018
Wiki-Vote	7,115	7,066	0.154	28.5	3.24	7	0.141	0.140
Electronic circuit S420	252	252	0.260	3.167	5.806	13	0.056	0.044
Electronic circuit S208	122	122	0.250	3.098	4.928	11	0.059	0.058

GCC (Giant Connected Component) size, m_D fraction of dominating nodes, $< k >$ average degree, l average shortest path, d diameter, C average clustering degree and NC network centrality

Why Not Contradicting [Liu et al.] ?

- Liu et al. assumed
  - only driver node values can be directly controlled through external signals.
- Conversely, MDS approach assumed
  - each driver node can control its links individually.

  ⇒ a node with degree k is regarded as k driver nodes.



(a) Model by Liu et al. (b) MDS model

Applications to Analysis of Biological Networks

MDS for Analyzing Biological Networks

- Applying control to real cells is far from easy
- However, MDS may be useful to find important proteins, genes, and other molecules
- Analysis of PPI networks
  - [Milenkovic et al. PLoS One, 2011] (before our work)
  - [Wuchty, PNAS, 2014]
  - [Khuri & Wuchty, BMC Bioinformatics, 2015]
  - [Wang et al., BIBM 2014]
- Analysis of metabolic cancer networks
  - [Asgari et al., PLoS ONE, 2013]

Application to Analysis of PPI Networks

- Wuchty found that MDS is useful to find important proteins [Wuchty, PNAS 2014]
  - Proteins in MDS are enriched with essential, cancer-related, and virus-targeted genes.
  - These proteins are highly involved in regulatory functions, showing high enrichment in transcription factors and protein kinases, and participate in regulatory links, phosphorylation events, and genetic interactions.

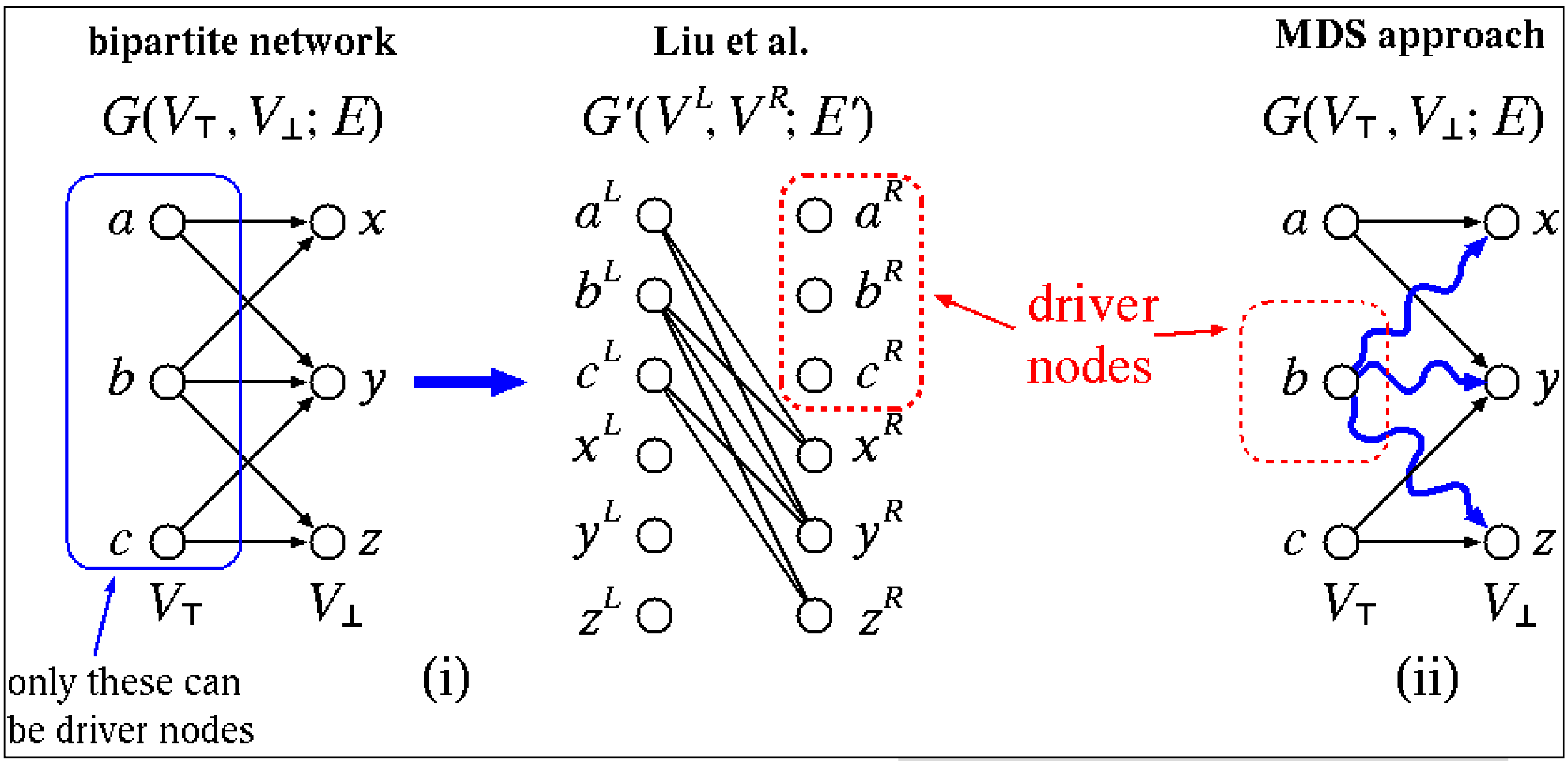
[Wuchty, PNAS 2014]

Extensions

Control of Bipartite Networks

Many real networks have bipartite structure (left/right nodes)

- Drug-target, researcher-paper, gene-disease
- Only left nodes can be driver nodes
- MDS approach needs much smaller number of driver nodes



[Nacher & Akutsu: Sci. Rep. 2013]

Results on Bipartite Networks

- New feature: Introduction of degree cutoff ($P(k)=0$ for $k > H$)
- For $\gamma_1<2$, the number of driver nodes is $O\left(\dfrac{n^{2-\gamma_1} m^{\gamma_1-1}}{H^{(2-\gamma_1)(\gamma_1-1)}}\right)$

Critical/Redundant Nodes in MDS

- We applied the concepts of critical/redundant nodes [Jia et al.: Nat. Comm. 2013] to MDS because MDS is not necessarily uniquely determined
  - Critical node: appears in every MDS
  - Redundant node: never appears in any MDS
- Critical nodes are expected to be more important than MDS

[Nacher & Akutsu: J. Comp. Net. 2015]

Robust MDS

- Robust MDS (RMDS): each node is dominated by at least C nodes (C=1 $\Rightarrow$ MDS)
  - Robust against deletion of arbitrarily C-1 edges
- Upper bound of RMDS size (for γ<2): $O\left(n^{1-\frac{(D-C+1)(2-\gamma)(\gamma-1)}{(D-C+1)(2-\gamma)+\gamma-1}}\right)$
  (D: minimum degree)
  - RMDS size corresponds to MDS size with minimum degree D-C+1

[Nacher & Akutsu: PRE, 2015]

Related Work by Molnar et al.

- Analysis of MDS size with degree cutoff [Sci. Rep. 2013]
- Analysis of MDS size with degree correlation [Sci. Rep. 2014]
- Damage-resilient dominating sets against random and targeted attacks [Sci. Rep. 2015]

Conclusion

Conclusion

- Establishment of a connection between MDS and structural controllability
- MDS size is small ($o(n)$) if $\gamma<2$

  $\Rightarrow$ Heterogeneous networks are not difficult to control
- This tendency was verified (to some extent) by computer simulation and database analysis
- Several extensions
  - Bipartite networks, Critical/Redundant nodes, Robust MDS
- MDS is useful for identifying important proteins in PPI networks

Future Work

- Development of a framework/theory which makes control of biological systems easy
- More rigorous theoretical analysis on MDS size (our analyses are based on a kind of mean-field approximation)
- More biological applications

Thank you !