Graph isomorphism determination

László Babai
2015 ACM Knuth Prize Winner

Graph Isomorphism (GI) problem is solved in quasipolynomial time: $2^{O((\log(n))^c)}$.

An isomorphism of graphs G and H is a bijection between the vertex sets of G and H that preserve the adjacency of two graphs.
Symmetry

- **Symmetry**: Invariance under a group of transformations (Wey.H)

- Key issues:
  - Invariance
  - Transformations
Graph Symmetry

- **Invariance of adjacency of vertices under the permutations on vertex set.**
- 'invariance' : the relation among the vertices
- 'transformation': is the permutation on vertex set.
Formalizations

- Permutation
- Automorphism
- Automorphism Group
- Structural Equivalent
- Orbit (trivial, non-trivial)
- Automorphism Partition
- Vertex Invariant
Why symmetry is so important?

- **Symmetry vs Complexity**
  - Taming complexity in nature and society are the major task of 21st century, and "complexity can then be characterized by lack of symmetry or 'symmetry breaking'" (F. Heylighen, 1996)

- **Evolution caused by Symmetry Breaking**
  - The universal evolution is caused by symmetry break, generating diversity and increasing complexity and energy (Mainzer K 2005; Quack 2003).
Outline

- What is symmetry?
- Application
  - Network Model
  - Network Measurement
  - Network Simplification
  - Shortest path index
  - Social network anonymization
Surprising! Real networks are symmetric!

Traditional Belief: 'almost all graphs are asymmetric'

TABLE I: Statistics of symmetry measures of some real networks. Tested symmetry measures includes $\alpha_G$, the automorphism group size of the real networks, to simplified the representation, we use $\log_{10} \alpha_G$; $C_G$ defined as ??; $\gamma_G$ defined as 1. All the networks is dealt as an undirected, unweighted graph without any loop and multi-edges. Basic information about the network (after processed), including $N$ (the number of vertex), $M$ (number of edges) and $z$ (average degree) are also showed.

<table>
<thead>
<tr>
<th>Network</th>
<th>$N$</th>
<th>$M$</th>
<th>$z$</th>
<th>$\log_{10} \alpha_G$</th>
<th>$\beta_G$</th>
<th>$\gamma_G(%)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>arXiv</td>
<td>27771</td>
<td>352285</td>
<td>25.37</td>
<td>333.2578</td>
<td>$1.006 \times 10^{-4}$</td>
<td>3.5074</td>
</tr>
<tr>
<td>InternetAS(c)</td>
<td>22443</td>
<td>45550</td>
<td>4.06</td>
<td>11346</td>
<td>$3.8784 \times 10^{-4}$</td>
<td>76.1</td>
</tr>
<tr>
<td>BioGrid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAC</td>
<td>5438</td>
<td>73054</td>
<td>13.43</td>
<td>57.7915</td>
<td>$5.118 \times 10^{-4}$</td>
<td>3.2739</td>
</tr>
<tr>
<td>MUS</td>
<td>219</td>
<td>400</td>
<td>3.65</td>
<td>126.9323</td>
<td>$4.687 \times 10^{-2}$</td>
<td>77.9817</td>
</tr>
<tr>
<td>HOM</td>
<td>7523</td>
<td>20029</td>
<td>5.32</td>
<td>935.0933</td>
<td>$4.808 \times 10^{-4}$</td>
<td>24.4685</td>
</tr>
<tr>
<td>DRO</td>
<td>7529</td>
<td>25196</td>
<td>6.69</td>
<td>624.3220</td>
<td>$4.2675 \times 10^{-4}$</td>
<td>21.36</td>
</tr>
<tr>
<td>CAE</td>
<td>2781</td>
<td>4350</td>
<td>3.13</td>
<td>829.6915</td>
<td>$1.941 \times 10^{-3}$</td>
<td>51.08</td>
</tr>
<tr>
<td>USPowerGrid</td>
<td>4942</td>
<td>3680</td>
<td>1.49</td>
<td>6169.8790</td>
<td>$9.744 \times 10^{-3}$</td>
<td>53.0864</td>
</tr>
</tbody>
</table>

Outline

- What is symmetry?
- Application
  - Network Model
  - Network Measurement
  - Network Simplification
  - Shortest path index
  - Social network anonymization
Emergence of symmetry in complex networks

Yanghua Xiao, 1 Momiao Xiong, 2, 3 Wei Wang, 1 and Hui Wang 4

1 Department of Computing and Information Technology, Fudan University, Shanghai 200433, People’s Republic of China
2 Theoretical Systems Biology Laboratory, School of Life Science, Fudan University, Shanghai 200433, People’s Republic of China
3 Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas 77225, USA
4 Business School, University of Shanghai for Science and Technology, Shanghai 200093, People’s Republic of China

(Received 21 September 2007; revised manuscript received 7 December 2007; published 11 June 2008)

Many real networks have been found to have a rich degree of symmetry, which is a universal structural property of complex networks, yet has been rarely studied so far. One of the fascinating problems related to symmetry is exploration of the origin of symmetry in real networks. For this purpose, we summarized the statistics of local symmetric motifs that contribute to local symmetry of networks. Analysis of these statistics shows that the symmetry of complex networks is a consequence of similar linkage pattern, which means that vertices with similar degrees tend to share common neighbors. An improved version of the Barabási-Albert model integrating similar linkage pattern successfully reproduces the symmetry of real networks, indicating that similar linkage pattern is the underlying ingredient that is responsible for the emergence of symmetry in complex networks.

DOI: 10.1103/PhysRevE.77.066108 PACS number(s): 89.75.Fb, 05.40.—a, 02.20.—a
Origination of symmetry in real networks

- Similar Linkage Pattern
  - *nodes having similar property such as degree, tend to have similar neighbors.*
  - Exact, vs non-exact

- (Generalized) Symmetry Bicliques

[Diagrams showing bicliques]

Symmetry in Complex Networks
**Model based on SLP**

- *Example*: when you first time join a social network, you not only link to stars (hubs), but also want to link to those accounts to whom most of your friends already in the network link.

- *Preferential attachment with similar linkage pattern*

\[
\Pi(v_i) = \begin{cases} 
\frac{\alpha k_i}{\sum_j k_j} + (1 - \alpha) \frac{1}{|V'_t(m)|} & \text{if } v_i \in V'_t(m), \\
\alpha \frac{k_i}{\sum_j k_j} & \text{if } v_i \notin V'_t(m),
\end{cases}
\]
With SLP, we can reproduce the network symmetry
Outline

- What is symmetry?
- Application
  - Network Model
  - Network Measurement
  - Network Simplification
  - Shortest path index
  - Social network privacy protection
Symmetry-based structure entropy of complex networks

Yang-Hua Xiao\textsuperscript{a,\ast}, Wen-Tao Wu\textsuperscript{a}, Hui Wang\textsuperscript{b}, Momiao Xiong\textsuperscript{c,d}, Wei Wang\textsuperscript{a}

\textsuperscript{a} Department of Computing and Information Technology, Fudan University, Shanghai 200433, PR China
\textsuperscript{b} Business school, University of Shanghai for Science and Technology, Shanghai 200043, PR China
\textsuperscript{c} Theoretical Systems Biology Lab, School of Life Science, Fudan University, Shanghai 200093, PR China
\textsuperscript{d} Human Genetics Center, University of Texas Health Science Center at Houston, Houston TX 77225, USA

Received 30 September 2007; received in revised form 20 December 2007
Available online 9 January 2008

Abstract

Precisely quantifying the heterogeneity or disorder of network systems is important and desired in studies of behaviors and functions of network systems. Although various degree-based entropies have been available to measure the heterogeneity of real networks, heterogeneity implicated in the structures of networks can not be precisely quantified yet. Hence, we propose a new structure entropy based on automorphism partition. Analysis of extreme cases shows that entropy based on automorphism partition can quantify the structural heterogeneity of networks more precisely than degree-based entropies. We also summarized symmetry and heterogeneity statistics of many real networks, finding that real networks are more heterogeneous in the view of automorphism partition than what have been depicted under the measurement of degree-based entropies; and that structural heterogeneity is strongly negatively correlated to symmetry of real networks.

© 2008 Elsevier B.V. All rights reserved.

Keywords: Structure entropy; Symmetry; Complex network
How to quantify the complexity of a graph?

- Observation:
  - Vertex with the same degree can be distinguished by many other metrics.
  - Vertex within a cell of automorphism partitioning cannot be distinguished by any structural metrics.

- Automorphism Partition is finer than Degree Partition.

Automorphism partitions of networks can naturally partition the vertex set into structurally equivalent cells.
## Heterogeneity of real networks

<table>
<thead>
<tr>
<th>Network</th>
<th>N</th>
<th>M</th>
<th>z</th>
<th>( \log \alpha_G )</th>
<th>( \beta_G )</th>
<th>( \gamma_G(%) )</th>
<th>NEDD</th>
<th>NERDD</th>
<th>NEAP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technique Network</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>USPowerGrid[31]</td>
<td>4942</td>
<td>6594</td>
<td>2.67</td>
<td>152.71</td>
<td>( 5.90 \times 10^{-4} )</td>
<td>16.7</td>
<td>0.20</td>
<td>0.25</td>
<td>0.98</td>
</tr>
<tr>
<td>InternetAS[^b]</td>
<td>2243</td>
<td>4550</td>
<td>4.06</td>
<td>11346</td>
<td>( 3.8784 \times 10^{-4} )</td>
<td>76.1</td>
<td>0.16</td>
<td>0.39</td>
<td>0.84</td>
</tr>
<tr>
<td><strong>Social Network</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>arXiv[^c]</td>
<td>27771</td>
<td>352285</td>
<td>25.37</td>
<td>333.26</td>
<td>( 1.01 \times 10^{-4} )</td>
<td>3.51</td>
<td>0.41</td>
<td>0.51</td>
<td>0.99</td>
</tr>
<tr>
<td>USAir97[^4]</td>
<td>332</td>
<td>2126</td>
<td>12.81</td>
<td>24.41</td>
<td>( 9.59 \times 10^{-3} )</td>
<td>26.20</td>
<td>0.539</td>
<td>0.68</td>
<td>0.95</td>
</tr>
<tr>
<td>PairsP[^5]</td>
<td>10617</td>
<td>63782</td>
<td>12.02</td>
<td>632.80</td>
<td>( 2.90 \times 10^{-4} )</td>
<td>24.32</td>
<td>0.32</td>
<td>0.47</td>
<td>0.97</td>
</tr>
<tr>
<td>foldoc[^6]</td>
<td>13356</td>
<td>91471</td>
<td>13.6974</td>
<td>17</td>
<td>( 2 \times 10^{-4} )</td>
<td>0.80</td>
<td>0.32</td>
<td>0.39</td>
<td>1</td>
</tr>
<tr>
<td>Erdos82[^7]</td>
<td>6927</td>
<td>11850</td>
<td>3.42</td>
<td>4222.5</td>
<td>( 1.6 \times 10^{-3} )</td>
<td>73.75</td>
<td>0.15</td>
<td>0.44</td>
<td>0.77</td>
</tr>
<tr>
<td><strong>Biological Network</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BioGrid-SAC[^8]</td>
<td>5438</td>
<td>73054</td>
<td>13.43</td>
<td>57.79</td>
<td>( 5.12 \times 10^{-4} )</td>
<td>3.2739</td>
<td>0.48</td>
<td>0.61</td>
<td>1.00</td>
</tr>
<tr>
<td>BioGrid-MUS[^8]</td>
<td>219</td>
<td>400</td>
<td>3.65</td>
<td>126.93</td>
<td>( 4.69 \times 10^{-2} )</td>
<td>77.98</td>
<td>0.28</td>
<td>0.47</td>
<td>0.64</td>
</tr>
<tr>
<td>BioGrid-HOM[^8]</td>
<td>7523</td>
<td>20029</td>
<td>5.32</td>
<td>935.09</td>
<td>( 4.81 \times 10^{-4} )</td>
<td>24.47</td>
<td>0.28</td>
<td>0.43</td>
<td>0.94</td>
</tr>
<tr>
<td>BioGrid-DRO[^8]</td>
<td>7529</td>
<td>25196</td>
<td>6.69</td>
<td>624.32</td>
<td>( 4.27 \times 10^{-4} )</td>
<td>21.36</td>
<td>0.30</td>
<td>0.45</td>
<td>0.96</td>
</tr>
<tr>
<td>BioGrid-CAE[^8]</td>
<td>2781</td>
<td>4350</td>
<td>3.13</td>
<td>829.69</td>
<td>( 1.94 \times 10^{-3} )</td>
<td>51.08</td>
<td>0.21</td>
<td>0.411</td>
<td>0.85</td>
</tr>
<tr>
<td>ppi[^9]</td>
<td>1870</td>
<td>2203</td>
<td>4.7132</td>
<td>518.6</td>
<td>( 2.7 \times 10^{-3} )</td>
<td>53.32</td>
<td>0.21</td>
<td>0.34</td>
<td>0.82</td>
</tr>
</tbody>
</table>

| **Theoretic Networks** |     |     |      |                       |                   |                  |      |       |      |
| Star Graph*         | 2000| 1999 | 1.99 | 5732.2                | \( 0.9962 \)      | 99.95            | 5.65 \times 10^{-4} | 0.09 | 5.65 \times 10^{-4} |
| BA (1)              | 2010| 2000 | 1.99 | 282.09                | \( 1.90 \times 10^{-3} \) | 56.37            | 0.17 | 0.30  | 0.91 |
| BA (2)              | 2010| 400  | 3.98 | 0.60                  | \( 1.40 \times 10^{-3} \) | 0.2             | 0.24 | 0.35  | 1    |
| BA (3)              | 2010| 600  | 5.97 | 0                     | \( 1.35 \times 10^{-3} \) | 0               | 0.28 | 0.39  | 1    |
| BA (4)              | 2010| 8000 | 7.96 | 0                     | \( 1.35 \times 10^{-3} \) | 0               | 0.31 | 0.43  | 1    |
| ER (1)              | 2000| 2081 | 2.08 | 507.97                | \( 2.4 \times 10^{-3} \) | 34              | 0.225| 0.228 | 0.89 |
| ER (2)              | 2000| 4002 | 4    | 51.33                 | \( 1.4 \times 10^{-3} \) | 2.65            | 0.276| 0.274 | 0.99 |
| ER (3)              | 2000| 5923 | 5.90 | 2.07                  | \( 1.36 \times 10^{-3} \) | 0.25            | 0.30 | 0.30  | 1    |
| ER (4)              | 2000| 8137 | 8.14 | 0                     | \( 1.36 \times 10^{-3} \) | 0               | 0.32 | 0.32  | 1    |

[^a]: All the networks are preprocessed as an undirected, unweighted graphs without any self-loops and multi-edges. Note that the whole

*From the symmetry perspective, the graph exhibits completely different complexity.*

Symmetry in Complex Networks
Outline

- What is symmetry?
- Application
  - Network Model
  - Network Measurement
  - Network Simplification
  - Shortest path index
  - Social network privacy protection
Network Quotients: structural skeletons of complex networks

Comments from one of PRE referee: ‘This manuscript makes a substantial contribution to the topic of finding the "structural skeleton" of complex networks, a topic of considerable interest in the network science literature.’

Network quotients: Structural skeletons of complex systems

Yanghua Xiao, Ben D. MacArthur, Hui Wang, Momiao Xiong, and Wei Wang

1 Department of Computing and Information Technology, Fudan University, Shanghai 200433, People’s Republic of China
2 School of Mathematics, University of Southampton, Southampton, SO17 1BJ, United Kingdom
3 Business School, University of Shanghai for Science and Technology, Shanghai 200093, People’s Republic of China
4 Theoretical Systems Biology Lab, School of Life Science, Fudan University, Shanghai 200433, People’s Republic of China
5 Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas 77225, USA

(Received 17 March 2008)

A defining feature of many large empirical networks is their intrinsic complexity. However, many networks also contain a large degree of structural repetition. An immediate question then arises: can we characterize essential network complexity while excluding structural redundancy? In this article we utilize inherent network symmetry to collapse all redundant information from a network, resulting in a coarse graining which we show to carry the essential structural information of the “parent” network. In the context of algebraic combinatorics, this coarse-graining is known as the “quotient.” We systematically explore the theoretical properties of network quotients and summarize key statistics of a variety of “real-world” quotients with respect to those of their parent networks. In particular, we find that quotients can be substantially smaller than their parent networks yet typically preserve various key functional properties such as complexity (heterogeneity and hubs vertices) and communication (diameter and mean geodesic distance), suggesting that quotients constitute the essential structural skeleton of their parent network. We summarize with a discussion of potential uses of quotients in analysis of biological regulatory networks and ways in which using quotients can reduce the computational complexity of network algorithms.

DOI: XXXX PACS number(s): 89.75.Fb, 05.40.--a, 02.20.--a
Motivation

Let’s make it simple!

Find the skeleton of complex network!
Network Quotient

- Quotient: $G/\text{Aut}(G)$
- Simplified quotient: *s quotient*
  - Coarse-graining each orbit into a vertex, preserving the adjacency between orbits
Quotients of real graphs

- Quotient graph is significantly smaller than its real graphs
- Many symmetric motifs contribute to the simplication
### Structural skeleton

- Mean geodesic distance: \( m \)
- Diameter: \( D \)
- Clustering coefficient \( C \)

<table>
<thead>
<tr>
<th>Network</th>
<th>( N )</th>
<th>( N_s )</th>
<th>( N_s/N )</th>
<th>( M )</th>
<th>( M_s )</th>
<th>( M_s/M )</th>
<th>( z )</th>
<th>( Z )</th>
<th>( Z_s )</th>
<th>( r )</th>
<th>( r_s )</th>
<th>( m )</th>
<th>( m_s )</th>
<th>( D )</th>
<th>( D_s )</th>
<th>( C )</th>
<th>( C_s )</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPI [31]</td>
<td>1458</td>
<td>1019</td>
<td>69.89%</td>
<td>1948</td>
<td>1469</td>
<td>75.41%</td>
<td>1.09</td>
<td>2.67</td>
<td>2.88</td>
<td>-0.21</td>
<td>-0.06</td>
<td>6.80</td>
<td>6.68</td>
<td>19</td>
<td>19</td>
<td>0.07</td>
<td>0.07</td>
</tr>
<tr>
<td>Yeast [32]</td>
<td>2224</td>
<td>1843</td>
<td>82.87%</td>
<td>6609</td>
<td>6133</td>
<td>92.80%</td>
<td>1.25</td>
<td>5.94</td>
<td>6.66</td>
<td>-0.11</td>
<td>-0.04</td>
<td>4.37</td>
<td>4.21</td>
<td>11</td>
<td>11</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>Homo [33]</td>
<td>7020</td>
<td>6066</td>
<td>86.41%</td>
<td>19811</td>
<td>18575</td>
<td>93.76%</td>
<td>1.30</td>
<td>5.64</td>
<td>6.12</td>
<td>-0.06</td>
<td>-0.01</td>
<td>4.86</td>
<td>4.77</td>
<td>14</td>
<td>14</td>
<td>0.10</td>
<td>0.11</td>
</tr>
<tr>
<td>California [34]</td>
<td>5925</td>
<td>4009</td>
<td>67.66%</td>
<td>15770</td>
<td>12882</td>
<td>81.69%</td>
<td>1.51</td>
<td>5.32</td>
<td>6.43</td>
<td>-0.23</td>
<td>-0.18</td>
<td>5.02</td>
<td>4.66</td>
<td>13</td>
<td>13</td>
<td>0.08</td>
<td>0.09</td>
</tr>
<tr>
<td>Epa [35]</td>
<td>4253</td>
<td>2212</td>
<td>52.01%</td>
<td>8897</td>
<td>6545</td>
<td>73.56%</td>
<td>1.15</td>
<td>4.18</td>
<td>5.92</td>
<td>-0.30</td>
<td>-0.16</td>
<td>4.50</td>
<td>4.11</td>
<td>10</td>
<td>10</td>
<td>0.07</td>
<td>0.10</td>
</tr>
<tr>
<td>DutchElite [36]</td>
<td>3621</td>
<td>1907</td>
<td>52.67%</td>
<td>4310</td>
<td>2576</td>
<td>59.77%</td>
<td>1.01</td>
<td>2.38</td>
<td>2.70</td>
<td>-0.24</td>
<td>-0.04</td>
<td>8.56</td>
<td>7.71</td>
<td>22</td>
<td>22</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Erdos02 [37]</td>
<td>6927</td>
<td>2365</td>
<td>34.14%</td>
<td>11850</td>
<td>7034</td>
<td>59.36%</td>
<td>1.06</td>
<td>3.42</td>
<td>5.95</td>
<td>-0.12</td>
<td>-0.08</td>
<td>3.78</td>
<td>3.41</td>
<td>4</td>
<td>4</td>
<td>0.12</td>
<td>0.29</td>
</tr>
<tr>
<td>Eva [38]</td>
<td>4475</td>
<td>898</td>
<td>20.07%</td>
<td>4652</td>
<td>1056</td>
<td>22.70%</td>
<td>1.01</td>
<td>2.08</td>
<td>2.35</td>
<td>-0.19</td>
<td>0.00</td>
<td>7.53</td>
<td>7.43</td>
<td>18</td>
<td>18</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>Geom [39]</td>
<td>3621</td>
<td>2803</td>
<td>77.41%</td>
<td>9461</td>
<td>7346</td>
<td>77.65%</td>
<td>2.59</td>
<td>5.23</td>
<td>5.24</td>
<td>0.17</td>
<td>0.19</td>
<td>5.71</td>
<td>5.15</td>
<td>14</td>
<td>14</td>
<td>0.54</td>
<td>0.43</td>
</tr>
<tr>
<td>P-fei1738 [40]</td>
<td>1738</td>
<td>1176</td>
<td>67.66%</td>
<td>1876</td>
<td>1312</td>
<td>69.94%</td>
<td>1.00</td>
<td>2.16</td>
<td>2.23</td>
<td>-0.27</td>
<td>0.07</td>
<td>10.22</td>
<td>10.39</td>
<td>29</td>
<td>29</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>InternetAs [41]</td>
<td>22442</td>
<td>11392</td>
<td>50.76%</td>
<td>45550</td>
<td>29564</td>
<td>64.90%</td>
<td>1.45</td>
<td>4.06</td>
<td>5.19</td>
<td>-0.20</td>
<td>-0.19</td>
<td>3.86</td>
<td>3.86</td>
<td>10</td>
<td>10</td>
<td>0.22</td>
<td>0.20</td>
</tr>
</tbody>
</table>

**Graph quotient preserves most structural properties of the original graph.**
Outline

- What is symmetry?
- Application
  - Network Model
  - Network Measurement
  - Network Simplification
  - Shortest path index
  - Social network privacy protection
Shortest Path Index

Motivation
- Shortest path trees can be used as index to answer shortest distance query
- $O(n^2)$ space complexity, unacceptable to big graphs

How to reduce the storage cost of BFS_trees?

Yanghua Xiao, Wentao Wu, JianPei, Wei Wang, Zhenying He, Efficiently Indexing Shortest Path by Exploiting Symmetry in Graphs, EDBT 2009, comment from referees: “A pioneer work”
BFS-trees mapping to each other

Many shortest path trees can be mapped to each other by automorphisms.

Figure 3: BFS-trees
Our Storage Solution

- Instead of storing all shortest path trees, we store a single shortest path tree for each orbit, and corresponding automorphisms.

- Automorphism: $g_1 = (v_1; v_2)$, $g_2 = (v_5; v_6)(v_7; v_9)(v_8; v_{10})$, $g_3 = (v_7; v_8)$, and $g_4 = (v_9; v_{10})$.

<table>
<thead>
<tr>
<th>Orbit</th>
<th>Base vertex</th>
<th>Mirrored vertex and automorphism</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Delta_1$</td>
<td>$v_1$</td>
<td>$\langle v_2, g_1 \rangle$</td>
</tr>
<tr>
<td>$\Delta_4$</td>
<td>$v_5$</td>
<td>$\langle v_6, g_2 \rangle$</td>
</tr>
<tr>
<td>$\Delta_5$</td>
<td>$v_7$</td>
<td>$\langle v_8, g_3 \rangle, \langle v_9, g_2 \rangle, \langle v_{10}, g_3g_2 \rangle$</td>
</tr>
</tbody>
</table>
Compact BFS-trees

- Each shortest path tree can be compressed further by its quotient
Our solution can significantly reduce the index size and speedup the query answering on the compressed index structure.
Outline

- What is symmetry?
- Application
  - Network Model
  - Network Measurement
  - Network Simplification
  - Shortest path index
  - Social network anonymization
Naïve Anonymization and Structural Re-identification (SR)

- Naïve Anonymization: replacing name with id

- Structural Re-identification (SR)
  - Suppose the adversary knows some structural knowledge $P$: “Bob has at least 4 neighbors”, then only node 2 satisfies $P$ and Bob will be re-identified.

Wentao Wu, Yanghua Xiao et.al, K-Symmetry Model for Identity Anonymization in Social Networks, EDBT’2010, March 22-26, Lausanne, Switzerland
**K-symmetry model**

- *Automorphism partition* of a graph

```
1 2 3
\-\-\-
4 5
\-\-
6 7 8
```

Vertices with the same color belong to the same cell (*orbit*) of the automorphism partition.

\[ \mathcal{O} = \{\{1, 3\}, \{2\}, \{4, 5\}, \{6, 8\}, \{7\}\} \]

- Vertices within the same orbit *cannot* be distinguished with each other by *any* structural knowledge.
- If each orbit contains at least \( k \) vertices, then the probability that an individual could be re-identified under *any* structural knowledge is *at most* \( 1/k \).
Orbit Copying Operation

- We achieve K-Symmetry by introducing orbit copying operations
  - Copying vertices in an orbit as well as their adjacency enough times until each orbit contains at least $k$ vertices
How to Use the Anonymized Network

- **Graph Utility:** Scientists are usually interested in the network’s *statistical properties*, such as degree distribution, average shortest path length, and so on.

- **Observation:** The anonymized graph although is quite different from the original graph, but they share the same *graph backbone*.

- **Solution**
  - Step 1: recover the graph backbone
  - Step 2: sample other elements under the constraint derived from the prior knowledge about the original graphs, such as the number of vertices
Utility Results

Backbone based sampling solution can recover most statistic information of the original graph such as degree, path length, transitivity, network resilience.
Yanghua Xiao and Hua Dong, Li Jin, Wei Wang, Momiao Xiong., *Evolution of Structure of Metabolic Networks*, ICSB2007 (8th International Conference on Systems Biology 2007, Hiroshima, Japan)


Hua Dong and Yanghua Xiao, Wei Wang, Li Jin, Momiao Xiong, *Symmetry in Metabolic Networks*. *Journal of Computer Science and System Biology*.


Wentao Wu, Yanghua Xiao, Wei Wang, Zhenying He and Zhihui Wang, K-Symmetry Model for Identity Anonymization in Social Networks (pdf), 13th International Conference on Extending Database Technology (EDBT’10)
Our other works

Real graphs are very big
- How to efficiently and effectively manage and analyze these big graphs (even with billions of nodes)?
- Shortest distance query (VLDB2014), big graph systems (SIGMOD12), Overlapping community search (SIGMOD2013), Community search (SIGMOD2014), Big graph partitioning (ICDE2014)

Real graphs are semantic rich
- How to construct knowledge graphs (KG) and how to use them in search, recommendation, and inference?

Our team: [http://gdm.fudan.edu.cn](http://gdm.fudan.edu.cn)
Our systems and knowledge bases [http://kw.fudan.edu.cn](http://kw.fudan.edu.cn)
What is the foundation of data science?

- Data understanding: Enable machine to understand data without supervision
  - First, data understanding is the prerequisite of all other data processing techniques
  - Second, many other challenges (such heterogeneity, big scale) can be considered as an obstacle of machine data understanding
  - Third, free your brains instead of just free your hands

- Challenge
  - What is understanding? How to model machine’s understanding?
  - What is the limit of machine’s understanding of data? Will we humans be replaced by machines?
Thanks!