Inference on Network Data

James P. Ferry and J. Oren Bumgarner
Metron, Inc.

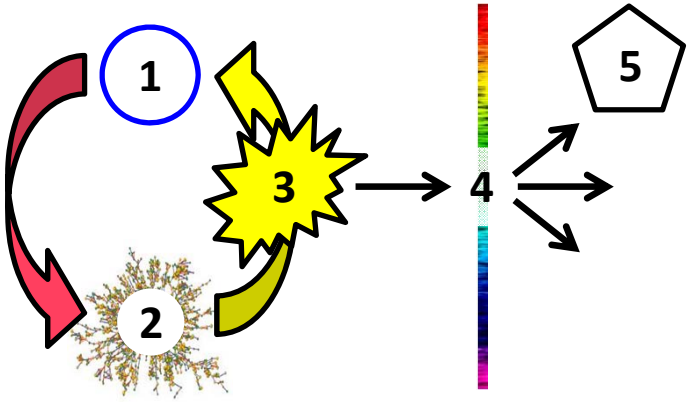
Graph Exploitation Symposium

August 9-10, 2011

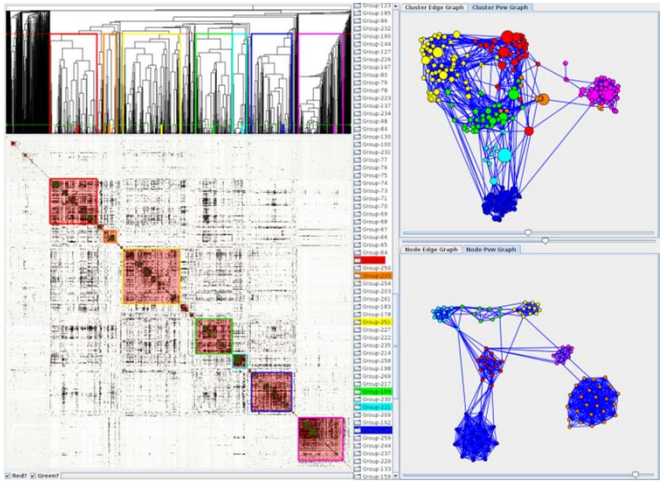
Lincoln Laboratory, Dedham, MA

Motivation and Outline

- Community detection problem
  - Many algorithms exist
  - But they make hard calls only (π)
- View in a data fusion paradigm →
  - State-to-measurement model
  - Inference in terms of probabilities



- Membership probability $p_v(C)$
  - Enhanced probability information for hard calls
- Co-membership probability $p^{\{v,w\}}$
  - Probabilistic summary of group structure
  - Fast approximation
- IGNITE
  - Network analysis tool based on $p^{\{v,w\}}$ →



Membership Probability $p_v(C)$

Probabilities that augment hard calls

Membership Probability

- Many algorithms produce a hard-call *partition* π of a network G
  - E.g., Infomap, RN, Louvain
- Let's augment this with the *probability* $p_v(C)$ that node v is in the group $C \in \pi$
- Planted partition model
  - $\Pr\left(G \middle| \pi, p_I, p_O\right) = p_I^{e_I}\left(1-p_I\right)^{\tilde{e}_I} p_O^{e_O}\left(1-p_O\right)^{\tilde{e}_O}$
  - Uniform distribution on p_I and p_O
  - Uniform distribution on $\log|\pi|$ and uniform on π given $|\pi|$
- Bayesian inversion
  - $\Pr\left(\pi \middle| G\right) \propto \dfrac{|\pi| \mathrm{B}\left(e_I+1, \tilde{e}_I+1\right) \mathrm{B}\left(e_O+1, \tilde{e}_O+1\right)}{S\left(n, |\pi|\right)}$
  - Given a partition π and a node v, compute the probabilities of the partitions formed by putting v in each group $C \in \pi$ or in its own group
  - Result: $p_v(C)$

$p_I =$ in-group edge probability

$e_I =$ # of in-group edges

$\tilde{e}_I =$ # of in-group non-edges

O quantities: between groups

Membership Probability

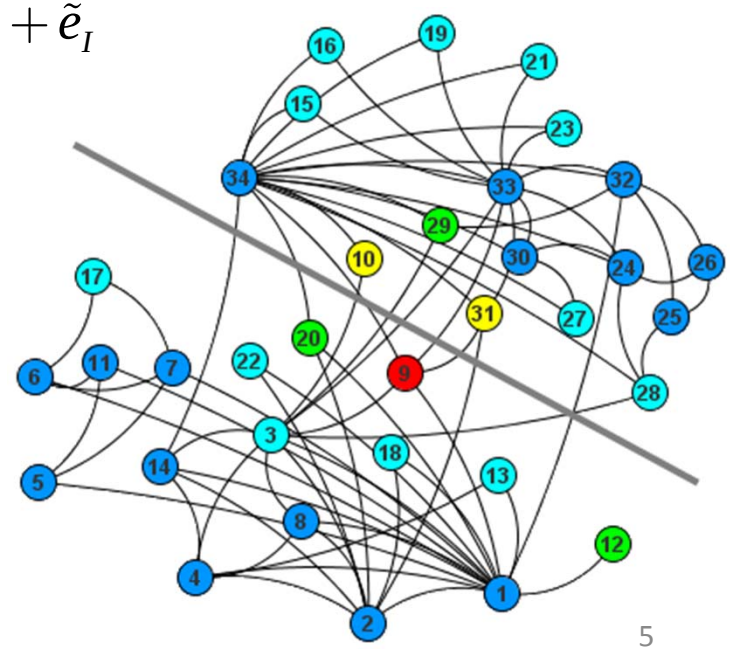
- Formula for $p_v(C)$:
  - $$p_v(C) \propto \alpha_C \frac{F(e_I, e_O, \delta_v(C)) F(\tilde{e}_I, \tilde{e}_O, \tilde{\delta}_v(C))}{F(h_I + 1, h_O + 1, \eta_v(C))} \approx \boxed{\alpha_C \gamma^{\delta_v(C)} \tilde{\gamma}^{\tilde{\delta}_v(C)}}$$

    Negative information

    Positive information
  - $\alpha_C =$ factor of $m/(1+1/m)^n$ if $C = \varnothing$
  - $\delta_v(C) =$ # of in-group edges when $v \to C$

    $\tilde{\delta}_v(C) =$ # of in-group non-edges when $v \to C$
  - $\gamma = \dfrac{e_I / h_I}{e_O / h_O}, \quad \tilde{\gamma} = \dfrac{\tilde{e}_I / h_I}{\tilde{e}_O / h_O}$ where $h_I = e_I + \tilde{e}_I$
- Example: Karate club data
  - Famous real-world dataset
  - $p_v(C_v)$ shown (probability of v being in its true group C_v, given G)
  - Red: 13%, yellow: 56%, green: 90%
  - Red node (9) often "misclassified"

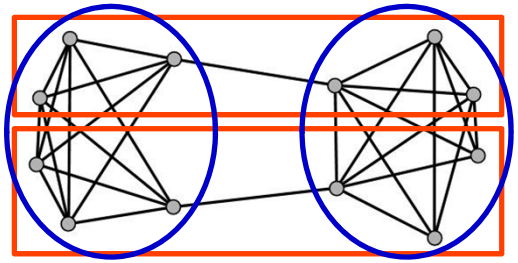
Karate club $p_v(C_v)$



Co-membership Probability $p^{\{v,w\}}$

Probabilistic alternative to hard calls

Bayesian Community Detection



- What makes <u>one</u> partition (π_{blue}) better than <u>another</u> (π_{red})?
- Usual answer: some network statistic
  - Network statistic = f (computed partition π^* given graph G)
- Bayesian decision theory answer:
  - Utility = f (computed partition π^* given true partition π)
  - Expected utility: $\mathbb{E}[U](\pi^*|G) \doteq \sum_{\pi} U(\pi^*|\pi)\Pr(\pi|G)$
- For a certain generic utility function:
  - $\mathbb{E}[U](\pi^*|G) = \sum_{C\in\pi^*}\sum_{\{v,w\}\subseteq C}\left(p^{\{v,w\}} - \theta\right)$, where $\theta = \dfrac{u_{GA} - u_{GD}}{(u_{GA} - u_{GD}) + (u_{BD} - u_{BA})}$
  - Depends only on pairwise co-membership probabilities $p^{\{v,w\}}$
- Therefore $p^{\{v,w\}}$ contains information needed for hard calls…
  - And much more (e.g., soft calls, distances, a basis for tracking)
  - Fast approximation for $p^{\{v,w\}}$ the basis for IGNITE tool

Exact Co-membership Probability

- Planted partition model: $(\pi,G) \sim \mathrm{H}(n,\mu,p_I,p_O)$
  - n = number of nodes
  - μ = mean co-membership probability (= 1 / # of groups)
  - p_I = probability of links within groups
  - p_O = probability of link between groups

$$\Pr(\pi|\mu) = \left(\mu^{-1}\right)_{|\pi|} \mu^n$$

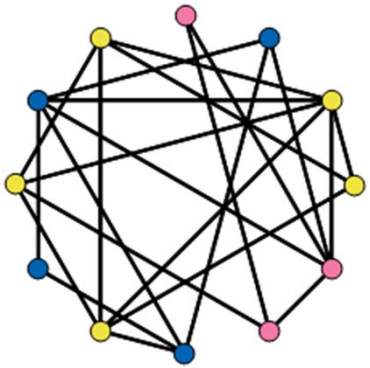
$$\Pr(G|\pi, p_I, p_O) = p_I^{e_I}(1-p_I)^{\tilde{e}_I}\, p_O^{e_O}(1-p_O)^{\tilde{e}_O}$$

$$\Pr(\pi|G, p_I, p_O, \mu) = \frac{\Pr(G|\pi, p_I, p_O)\Pr(\pi|\mu)}{\Pr(G|p_I, p_O, \mu)}$$

$$\Pr(\{v,w\}|G, p_I, p_O, \mu) = \sum_{\pi \succ \{v,w\}} \Pr(\pi|G, p_I, p_O, \mu)$$

$$p^{\{v,w\}} = \iiint \Pr(\{v,w\}|G, p_I, p_O, \mu)\Pr(p_I, p_O, \mu)\, dp_I\, dp_O\, d\mu$$

An instance (π,G) of $\mathrm{H}(12,1/3,0.8,0.1)$



Partition π = $\{\{1,6,8,10\}, \{2,3,7,9,11\},\{4,5,12\}\}$ (clock numbering)

- $p^{\{v,w\}}$ expensive : sum over vast number of partitions, cannot prune effectively

Approximate Co-membership Probability

- For fixed nodes v and w
  - Let H be the hypothesis that they are in the same group, and
  - E be some *evidence* derived from the graph G
- $p^{\{v,w\}} = \Pr(H|E)$: compute in terms of likelihood ratio
  - $\Pr(H|E) = \dfrac{\Lambda(E)}{\Lambda(E) + \bar{\mu}^{-1} - 1}$ where $\Lambda(E) \doteq \dfrac{\Pr(E|H)}{\Pr(E|\neg H)}$ and $\bar{\mu} = \Pr(H)$
  - Use uniform prior on $\log \mu$ uniform over $0 < p_O < p_I < 1$.
  - Then $\bar{\mu} = \int_{\mu_{MIN}}^{\mu_{MAX}} \mu \Pr(\mu) d\mu = \dfrac{\mu_{MAX} - \mu_{MIN}}{\log(\mu_{MAX}/\mu_{MIN})}$
  - Break down $\Pr(E|H)$ (and similarly for $\Pr(E|\neg H)$) as
    - $\Pr(E|H) = \int_{\mu_{MIN}}^{\mu_{MAX}} \Pr(E|H,\mu)\Pr(\mu) d\mu$
    - $\Pr(E|H,\mu) = 2\boxed{\int_0^1 \int_0^{p_I} \Pr(E|H, p_I, p_O, \mu) dp_O dp_I}$
- Need evidence E to be <u>powerful</u> but $\Pr(E|H, p_I, p_O, \mu)$ <u>tractable</u>

Graph Evidence

- Using E = entire graph G makes $\Pr\left(E \middle| H, p_I, p_O, \mu\right)$ a vast sum
- Let $E = (K_{vw}, K_{v\mathbf{x}}, K_{w\mathbf{x}})$ = all node pairs intersecting $\{v,w\}$
  - K_{yz}: random variable = 1 iff $\{y,z\}$ an edge, 0 else
  - $\mathbf{x}$ = all nodes other than v and w
  - E is the evidence most relevant to v and w being in the same group
- Posterior probability of evidence:

$$\Pr\left(E \middle| H, p_I, p_O, \mu\right) = p_I^{k_{vw}} \left(1 - p_I\right)^{1-k_{vw}} f\left(\delta, \phi_H\right), \quad \text{and}$$

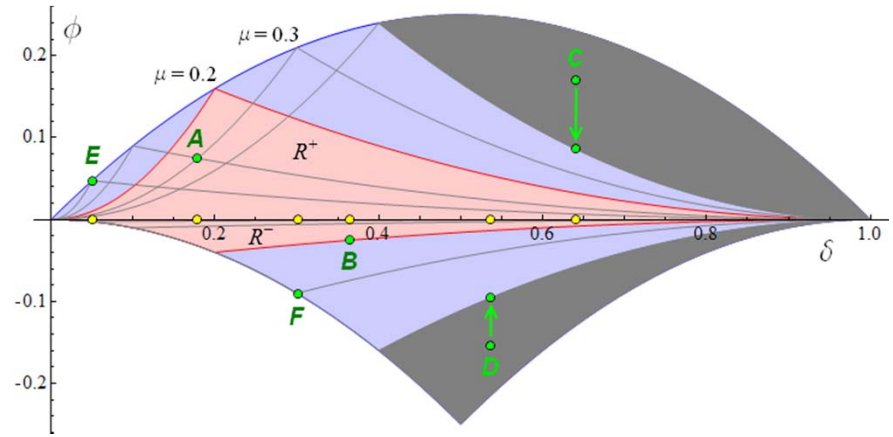
$$\Pr\left(E \middle| \neg H, p_I, p_O, \mu\right) = p_O^{k_{vw}} \left(1 - p_O\right)^{1-k_{vw}} f\left(\delta, \phi_{\neg H}\right) \quad \text{where}$$

$$f(\delta, \phi) = \left((1-\delta)^2 + \phi\right)^{n_0} \left(\delta(1-\delta) - \phi\right)^{n_1} \left(\delta^2 + \phi\right)^{n_2}$$

  - where n_0, n_1, n_2 = number of nodes adjacent to 0, 1, or 2 of $\{v,w\}$,
  - $\delta = \mathbb{E}\left[K_{vw}\right] = \mu p_I + (1-\mu) p_O$
  - $\phi_H = \mathrm{Cov}\left[K_{vx}, K_{wx} \middle| H\right] = \mu(1-\mu)\left(p_I - p_O\right)^2$
  - $\phi_{\neg H} = \mathrm{Cov}\left[K_{vx}, K_{wx} \middle| \neg H\right] = -\mu^2 \left(p_I - p_O\right)^2$

Integral in (δ,ϕ) space

- Two integrals:
  - Over $\phi > 0$ and over $\phi < 0$
- Integrals dominated by sharp peak in f
  - at global peak at (δ_p, ϕ_p)
  - or local peak at $(\delta_p, 0)$
  - …provided μ in range



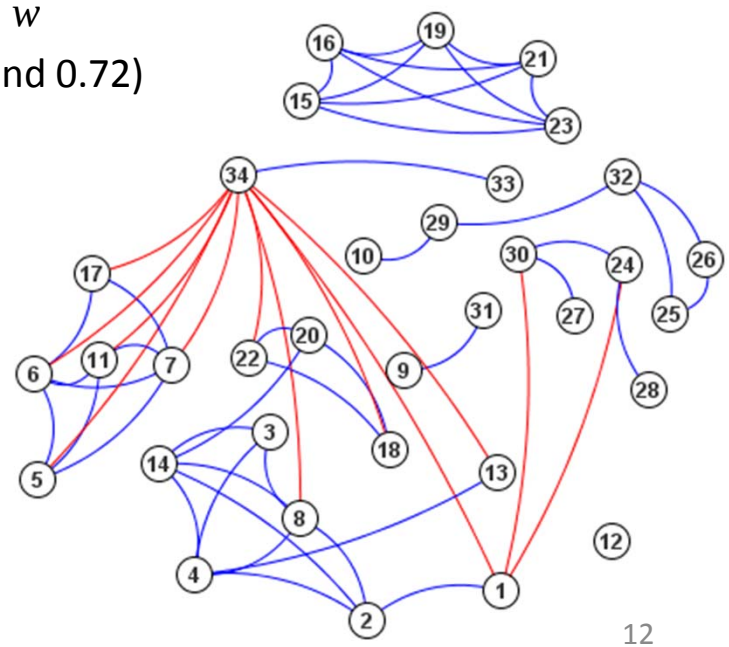
$$f(\delta,\phi) = \left((1-\delta)^2 + \phi\right)^{n_0} \left(\delta(1-\delta) - \phi\right)^{n_1} \left(\delta^2 + \phi\right)^{n_2}$$

- Peak location: $\delta_p = \dfrac{n_1 + 2n_2}{2(n_0 + n_1 + n_2)}, \quad \phi_p = \dfrac{4n_0 n_2 - n_1^2}{4(n_0 + n_1 + n_2)^2}$
- Global/local ratio: $\eta_p = \dfrac{f(\delta_p, \phi_p)}{f(\delta_p, 0)} = \left(1 + \dfrac{\phi_p}{(1-\delta_p)^2}\right)^{n_0} \left(1 - \dfrac{\phi_p}{\delta_p(1-\delta_p)}\right)^{n_1} \left(1 + \dfrac{\phi_p}{\delta_p^2}\right)^{n_2}$
- Likelihood ratio dominated by η_p or $1/\eta_p$, modified by
  - Contributions due to Jacobians and K_{vw} factor
  - Effect of peak being in μ domain only for limited range of μ

Co-membership Probability Formula

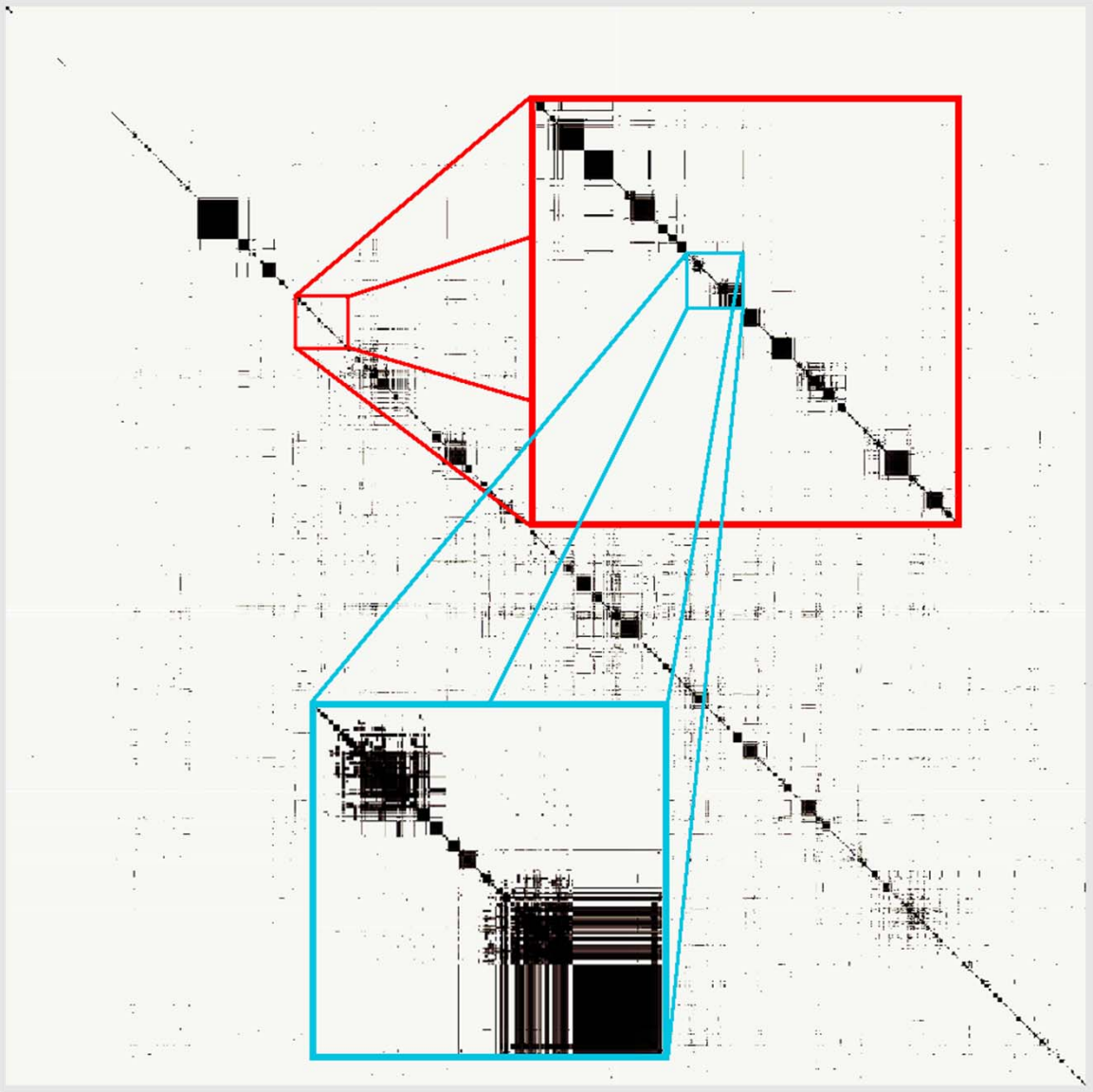
- Final approximation formula: $p^{\{v,w\}} = \dfrac{\Lambda}{\Lambda + \bar{\mu}^{-1} - 1}$
  - where $\bar{\mu} = (1/2 - 1/n)/\log(n/2)$ expresses prior distribution on group size,
  - and the likelihood ratio is split into two factors: $\Lambda = \tilde{\Lambda}\Lambda_{k_{vw}}$
    - $\tilde{\Lambda} = \eta_p$ if $\phi_p \geq 0$, $1/\eta_p$ else
      - Effect of presence/absence of edge on all node pairs containing *v* xor *w*
      - Multiples Λ by huge factor for $\phi_p >> 0$, divides by huge factor for $\phi_p << 0$
    - $\Lambda_0 = \min\left(0.7197, 0.46\delta_p^{-0.15}\right)$
      - Effect of absence of edge between *v* and *w*
      - Roughly constant (varies between 0.46 and 0.72)
      - Via theory + fit to numerical integration
    - $\Lambda_1 = \min\left(0.5605n + 1.598, \delta_p^{-0.7}\right)$
      - Effect of edge between *v* and *w*
      - Significant for low-degree nodes
  - Depends on n_0, n_1, n_2
    - Which in turn depends only on n_2 and the degrees of the nodes

Karate club $p^{\{v,w\}}$



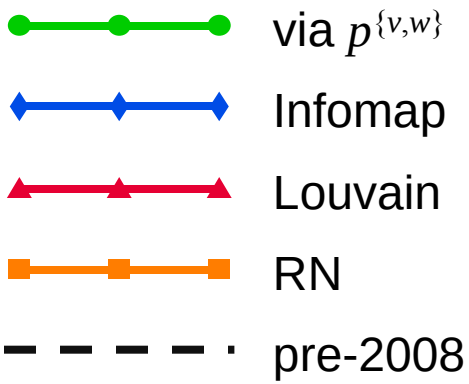
Visualizing $p^{\{v,w\}}$

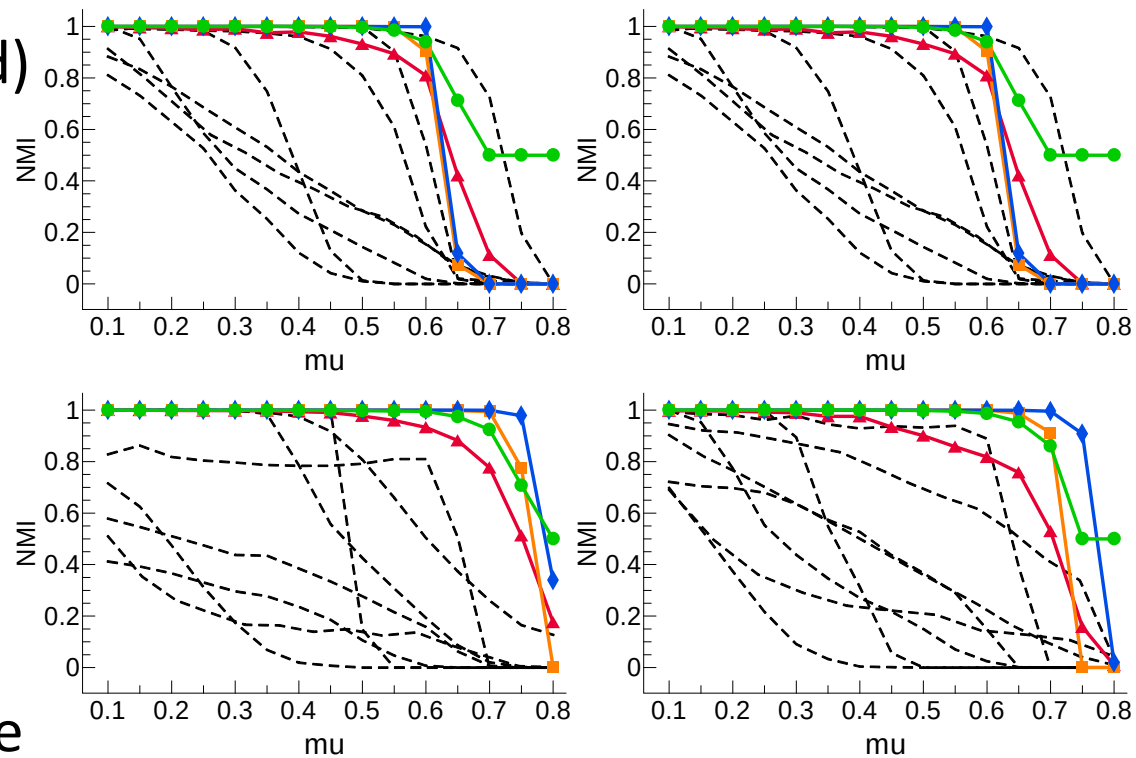
- Coloring edges ineffective
- Better to visualize matrix directly
- Reichardt and Bornholdt (2004)
  - But computed via expensive Monte Carlo methods
- Example: Enron email dataset
  - 36,692 nodes
  - 367,662 edges



Is it Accurate?

- Testing probability statements directly is difficult
- But $p^{\{v,w\}}$ has info to make hard calls (via expected utility)
- Allows us to compare to state-of-the-art hard-call methods
- Plot NMI vs. scenario difficulty (LFR testbed)





- Result: $p^{\{v,w\}}$ yields state-of-the-art hard-call performance

Is it Fast?

Dataset	Nodes	Links	Sum of n2	#(n2 > 0)	Triples	Groups
Caltech	769	16,656	1,231,412	186,722	14,120	19
Princeton	6,596	293,320	46,139,701	8,776,074	83,004	51
Georgetown	9,414	425,638	67,751,053	15,616,610	113,722	90
Oklahoma	17,425	892,528	194,235,901	47,202,925	239,162	233
UNC	18,163	766,800	140,796,299	47,576,619	191,482	167
Slashdot	82,168	504,230	74,983,589	49,450,449	104,330	5,209
Amazon	262,111	899,792	9,120,350	6,434,638	7,178	12,851
Livejournal	4,847,571	42,851,237	7,269,503,753	4,193,393,006		
Wb-edu	9,450,404	46,236,105	12,203,737,639	4,232,928,806		

Dataset	Infomap 10 trials	n2 1 proc	n2 -> null 1 proc	n2 8 proc	n2-> null 8 proc
Caltech	1.0	0.11	0.09	0.02	0.02
Princeton	45.4	4.5	3.9	0.67	0.55
Georgetown	77.5	7.2	6.4	1.2	0.93
Oklahoma	310.8	23.5	19.4	4.2	2.8
UNC	446.6	19.2	16.3	3.6	2.5
Slashdot	1553.4	14.5	12.6	3.0	1.9
Amazon	2075.8	2.4	2.2	0.59	0.44
Livejournal			1246.2		172.9
Wb-edu			1243.0		174.6

- Storage better than $O(n^2)$, but still significant
- Fast for dense (Facebook) networks
- Faster for sparse networks
- Parallelizable

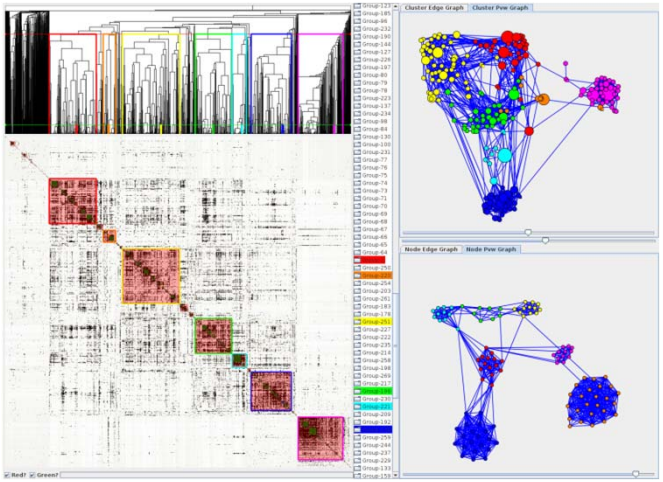
(All timings in seconds)

IGNITE

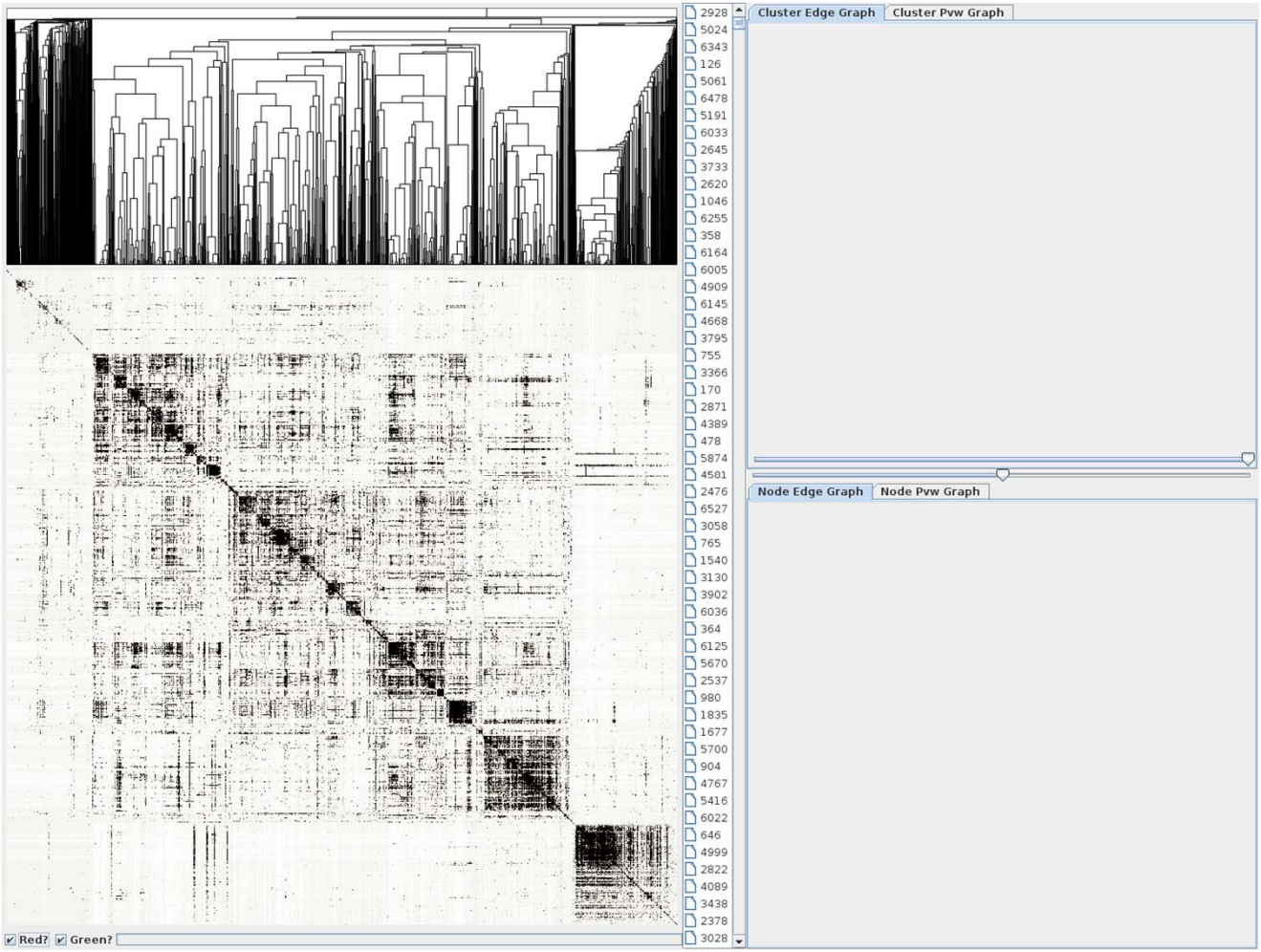
Inter-Group Network Inference and Tracking Engine

IGNITE

- Inter-Group Network Inference and Tracking Engine
- Uses $p^{\{v,w\}}$ as a basis for network analysis
- Uses distance metric $1 - p^{\{v,w\}}$ to hierarchically cluster nodes
  - Resulting dendrogram is entry point for defining
    - Groups (large sets of nodes)
    - Node clusters (small sets of nodes to treated a single nodes)
- Network visualized as $p^{\{v,w\}}$ matrix
- Groups selected for network view
  - Layout exploits $p^{\{v,w\}}$
  - Various drill-down and statistical tools
- Tracking technology
  - Subject of another talk (e.g. Fusion 2010)
  - Part of IGNITE codebase, but not integrated into tool

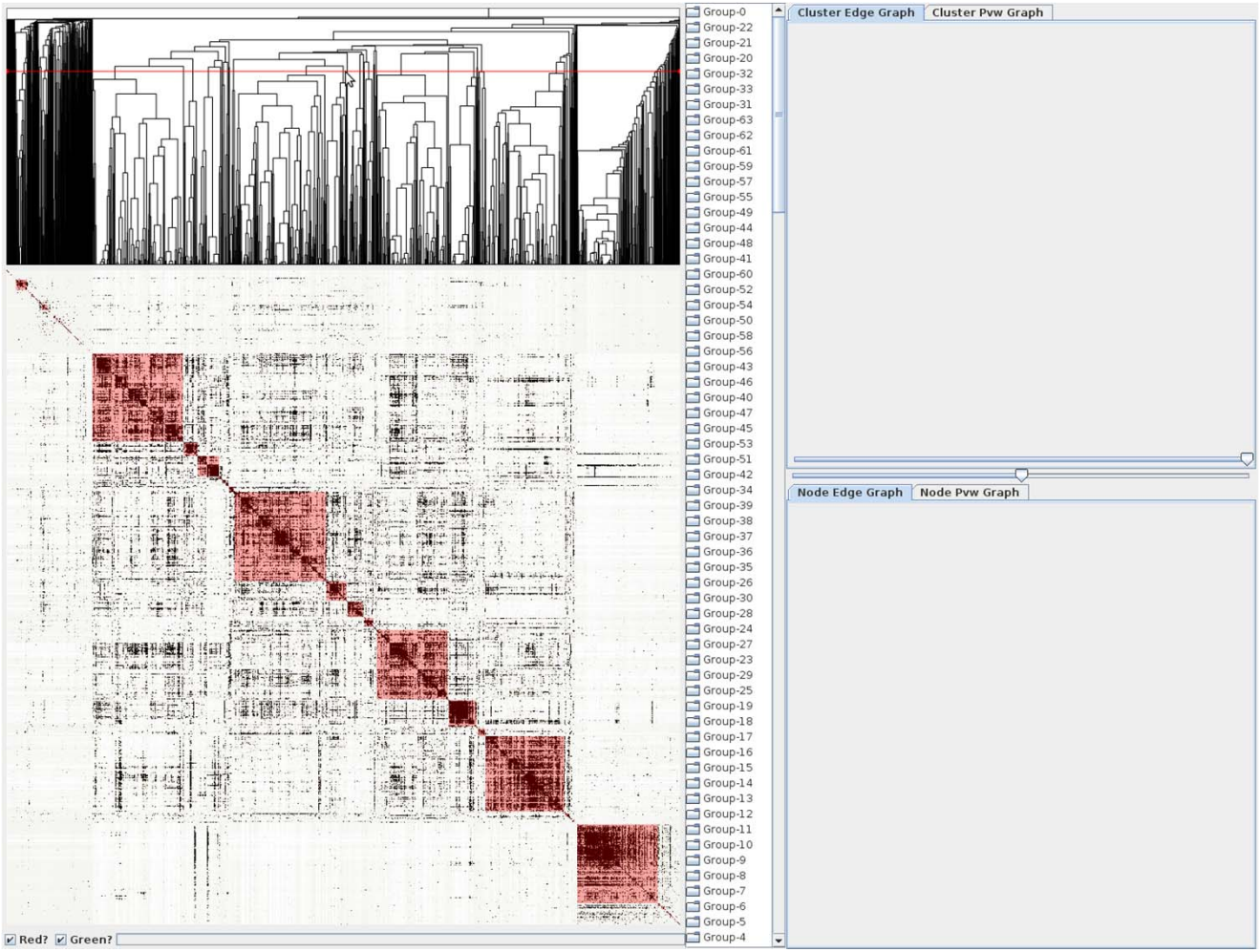


IGNITE Example



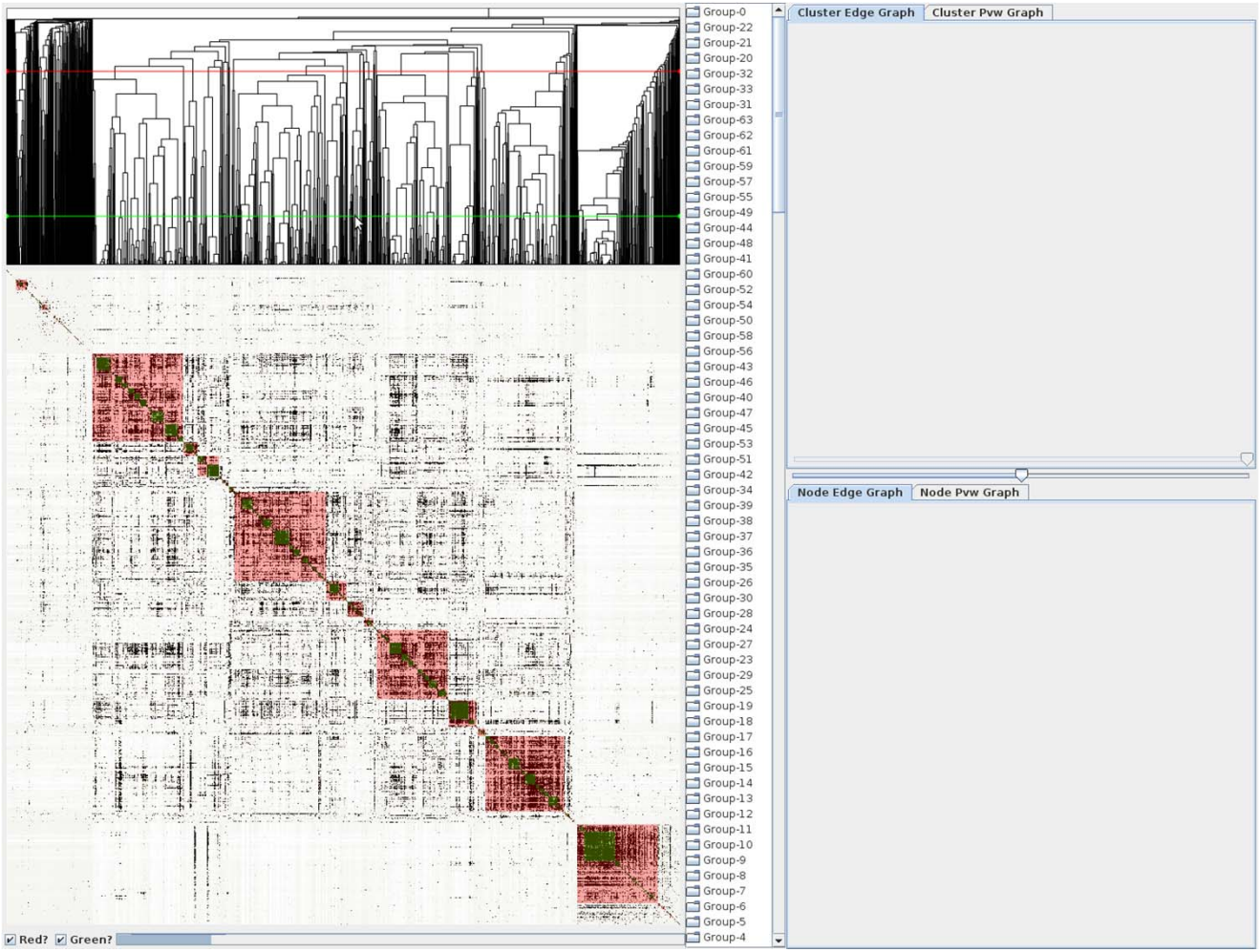
Tool at startup. Princeton Facebook network with 6596 nodes.

IGNITE Example



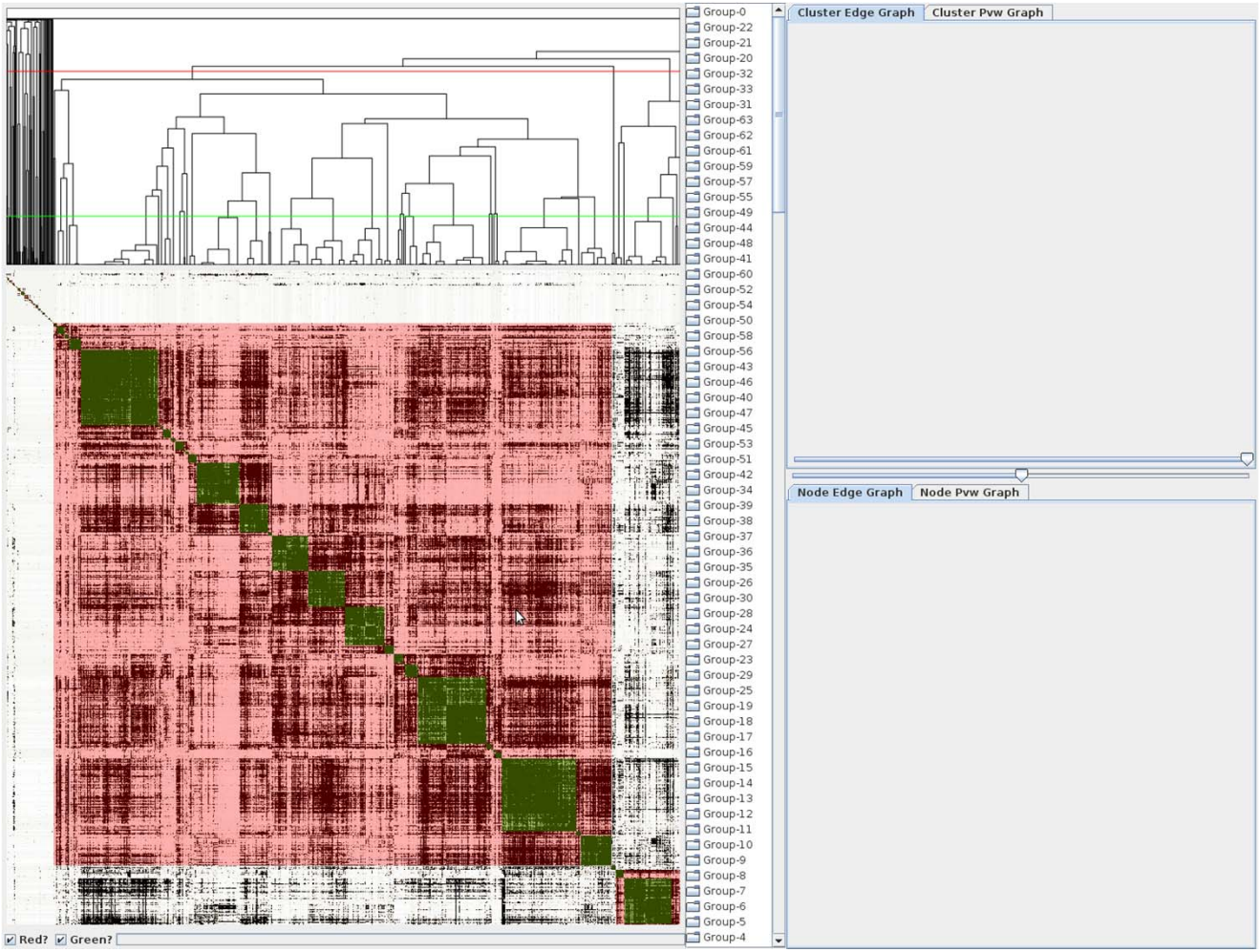
Click on the dendrogram to set an initial red partition.

IGNITE Example



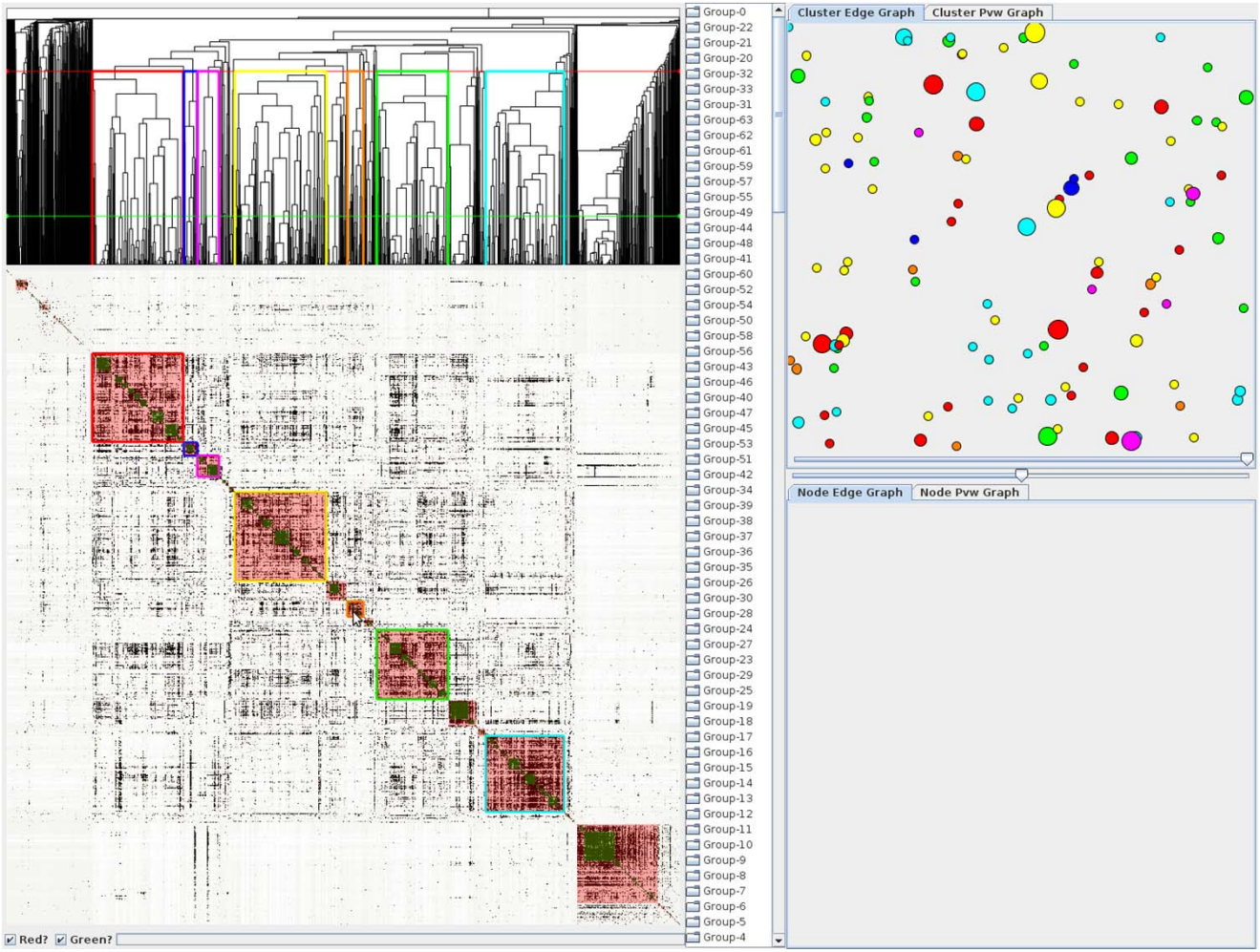
Click below the red to set an initial subpartition in green.

IGNITE Example



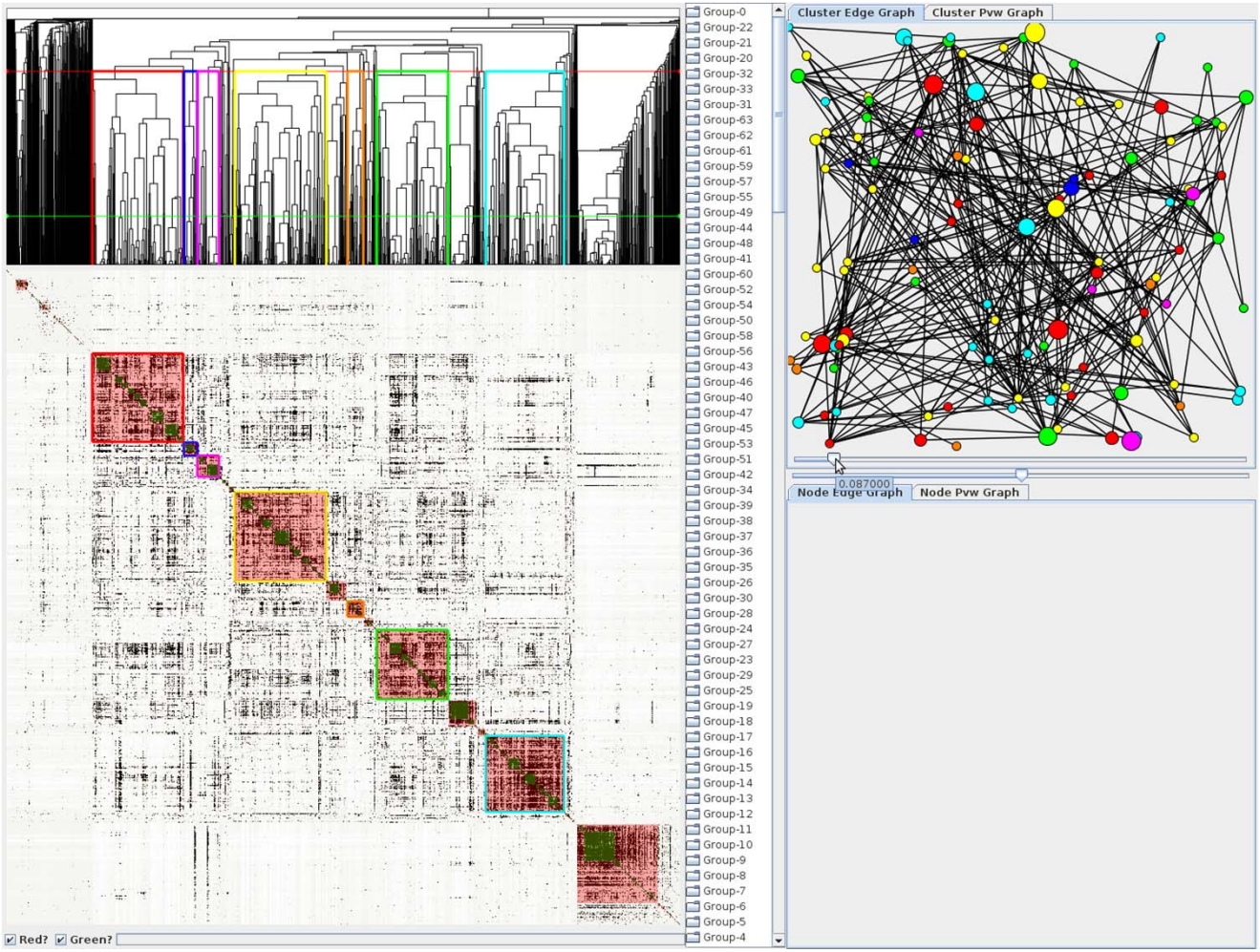
Click and drag to pan, zoom with mouse wheel to drill down. Dendrogram and $p^{\{v,w\}}$ plot are linked.

IGNITE Example



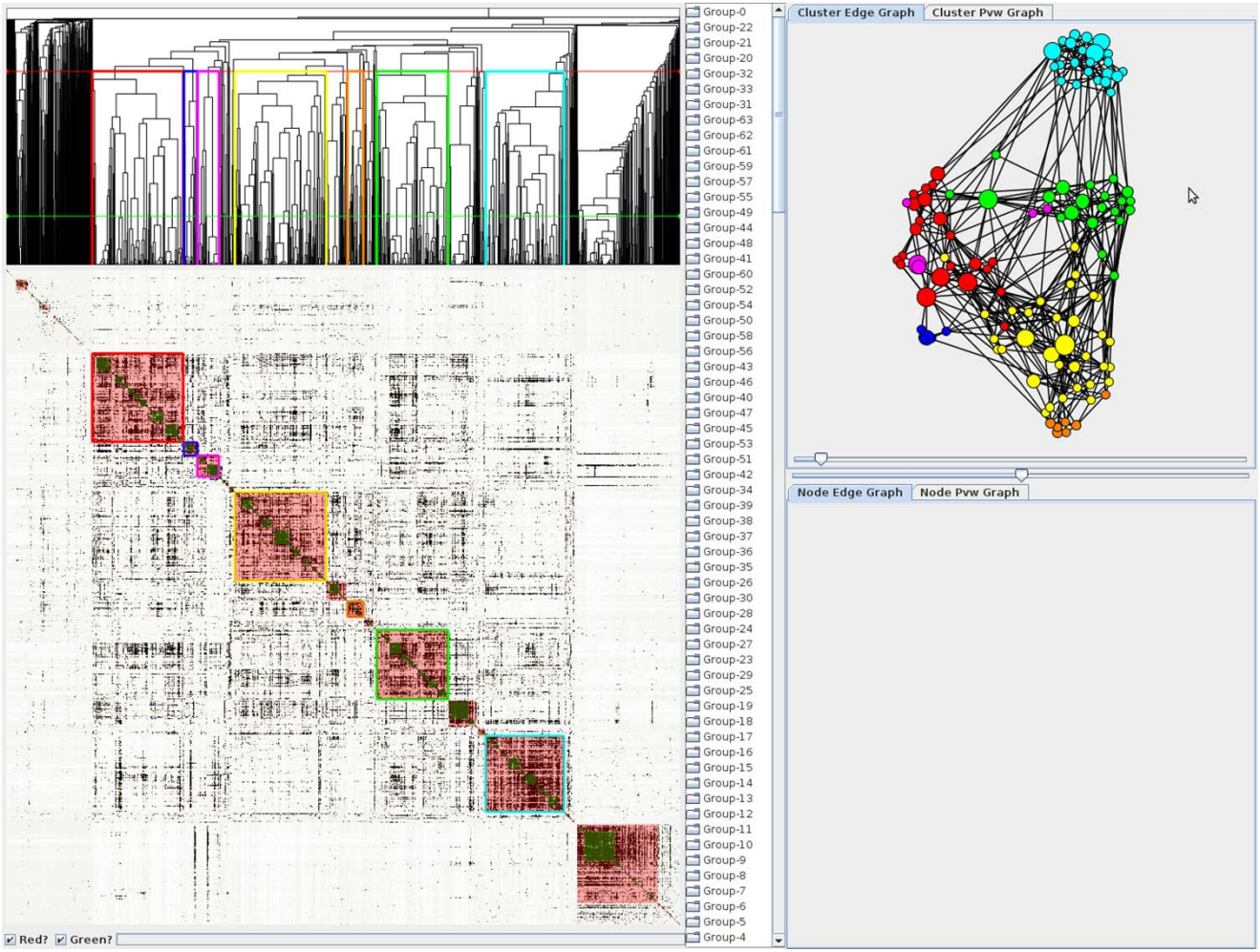
Clicking on a group in the $p^{\{v,w\}}$ plot at the red level highlights the group, and displays the subgroups in that color in the upper right hand panel.

IGNITE Example



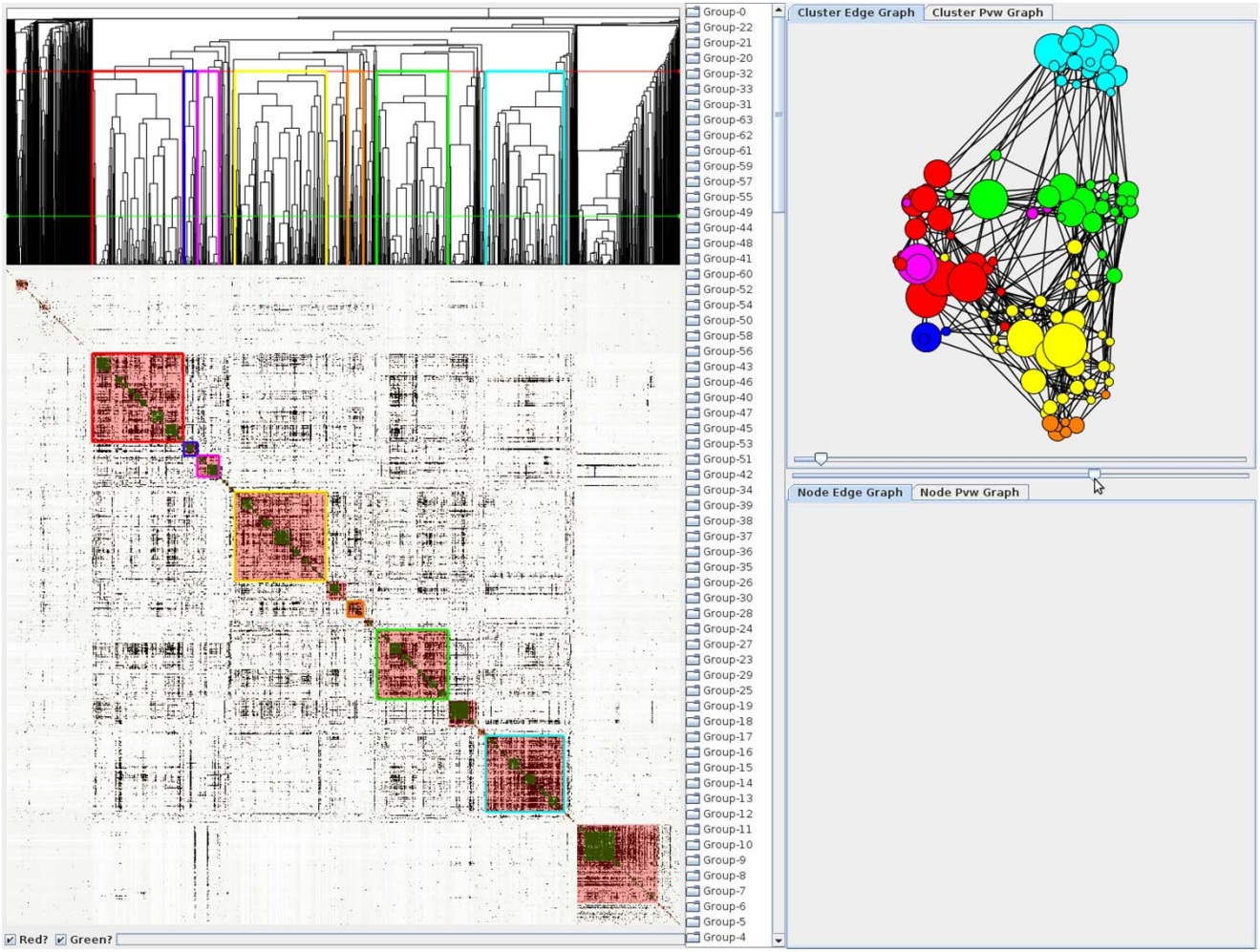
Drag the slider immediately below the graph to set the threshold for average link density between clusters.

IGNITE Example



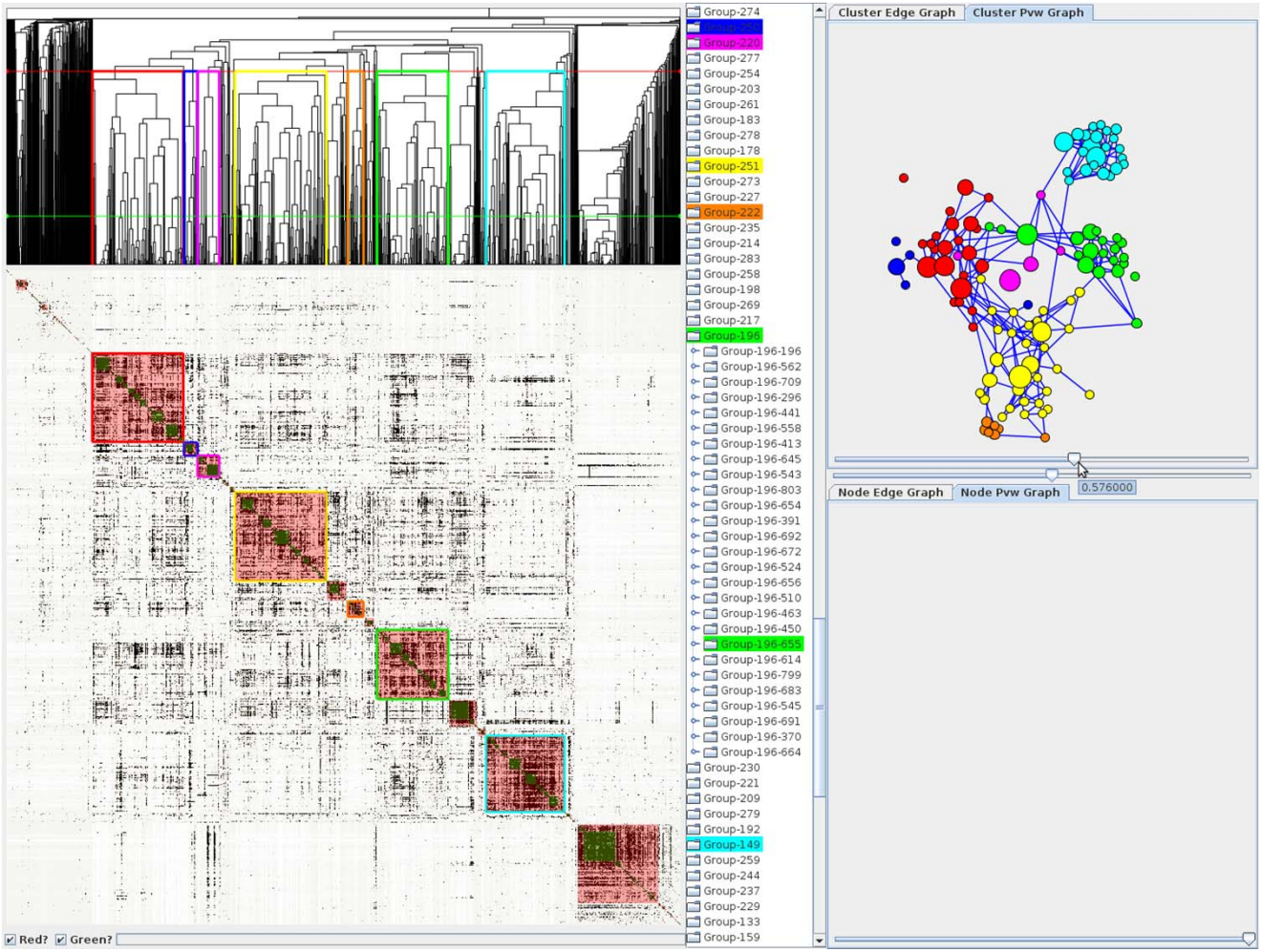
Right click in the panel and select the option to layout the graph according to the current link structure.

IGNITE Example



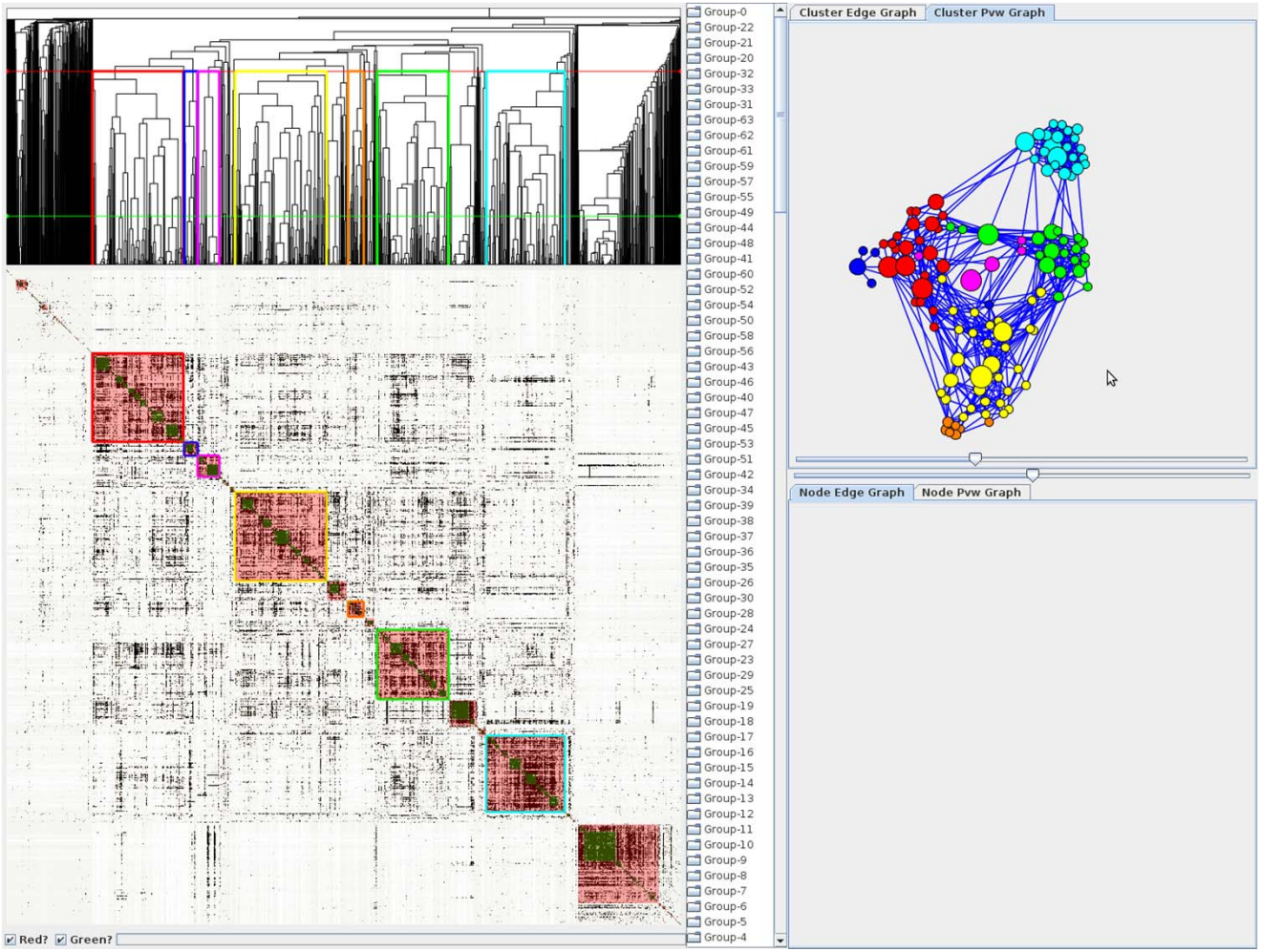
The second slider controls node size which is related to the number of nodes in the group.

IGNITE Example



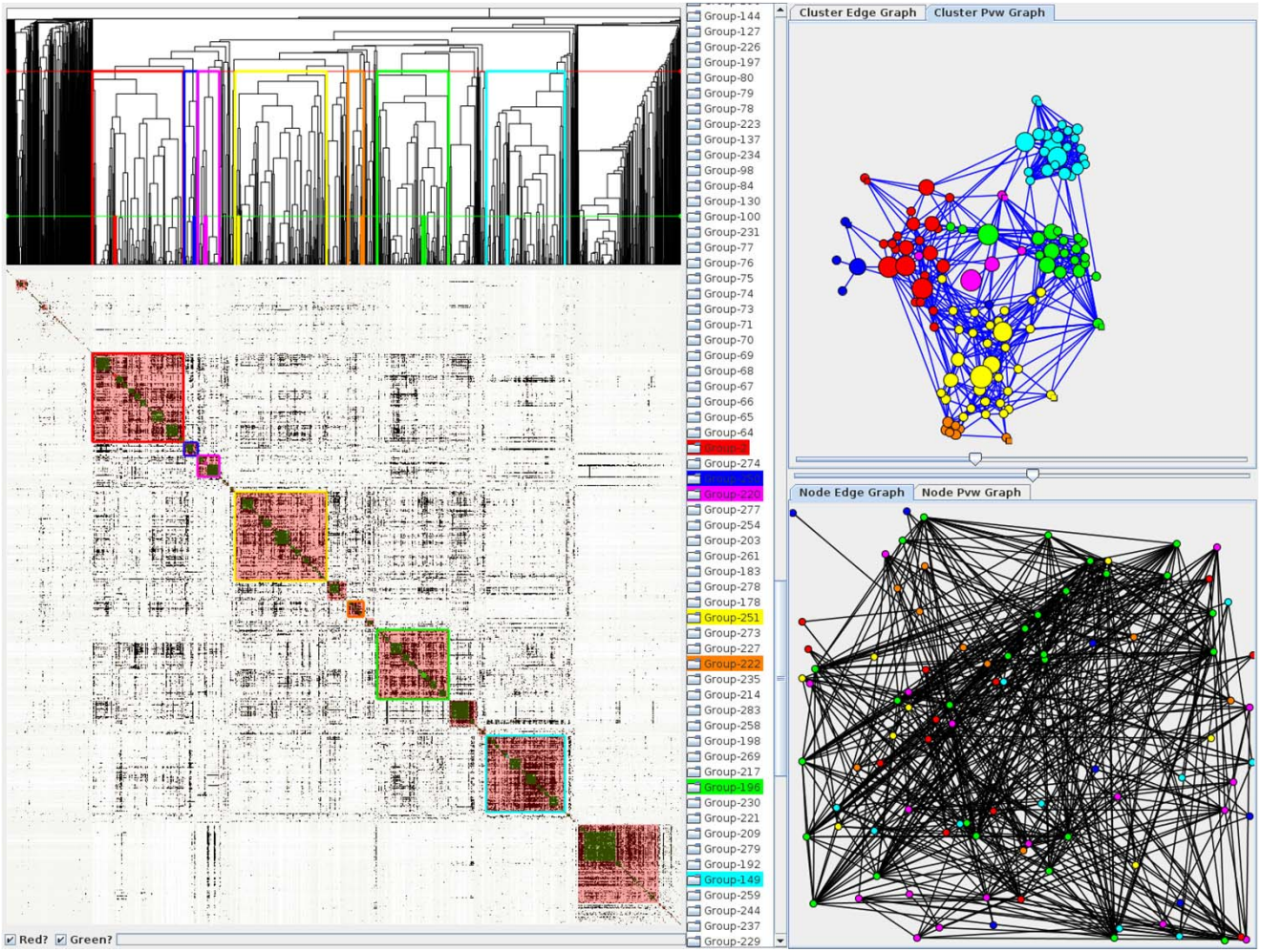
Switch to the Cluster Pvw Graph and drag the slider to adjust the threshold for mean Pvw required for a link.

IGNITE Example



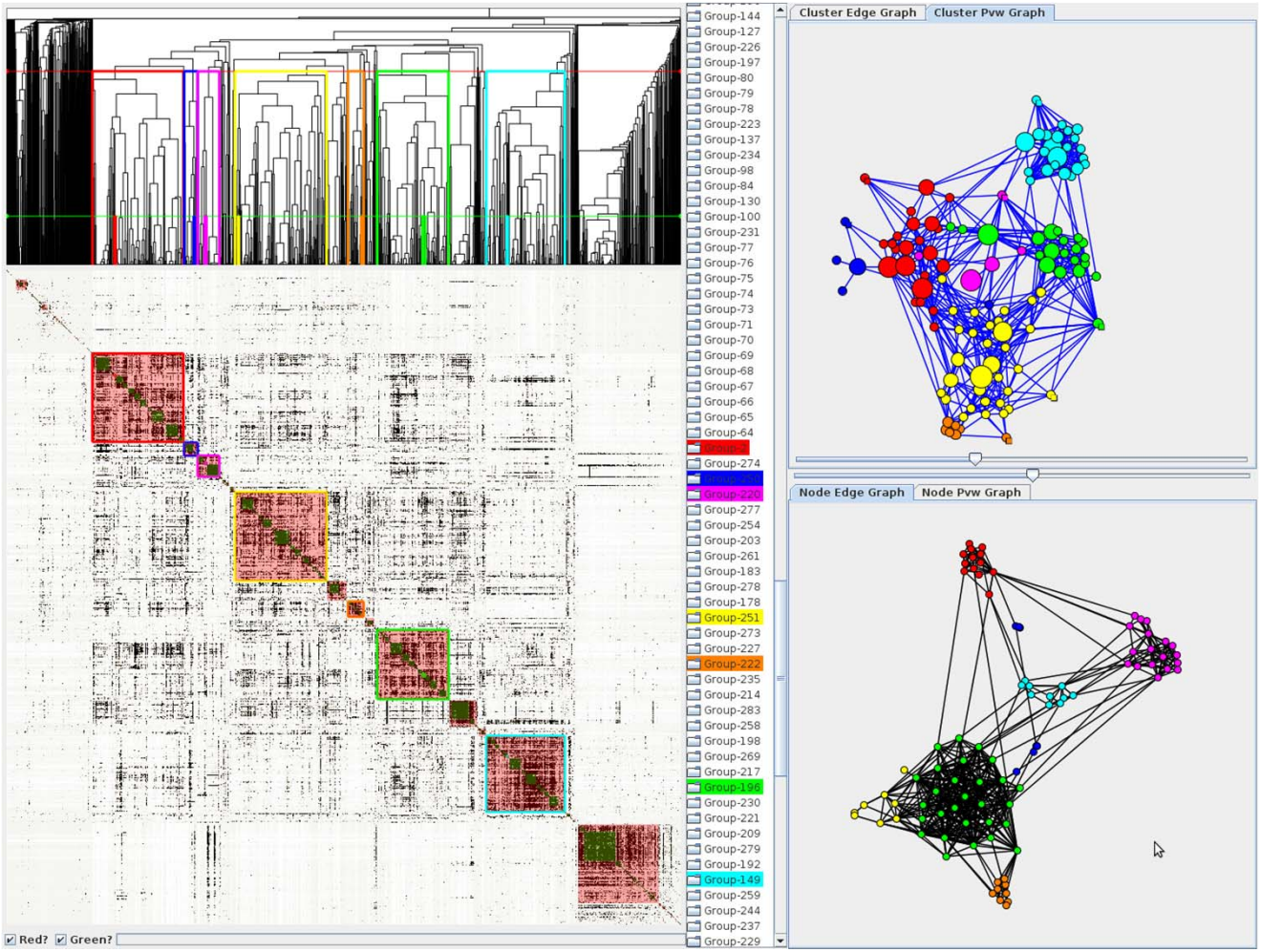
As before, the user can right click to run the layout according to the displayed links.

IGNITE Example



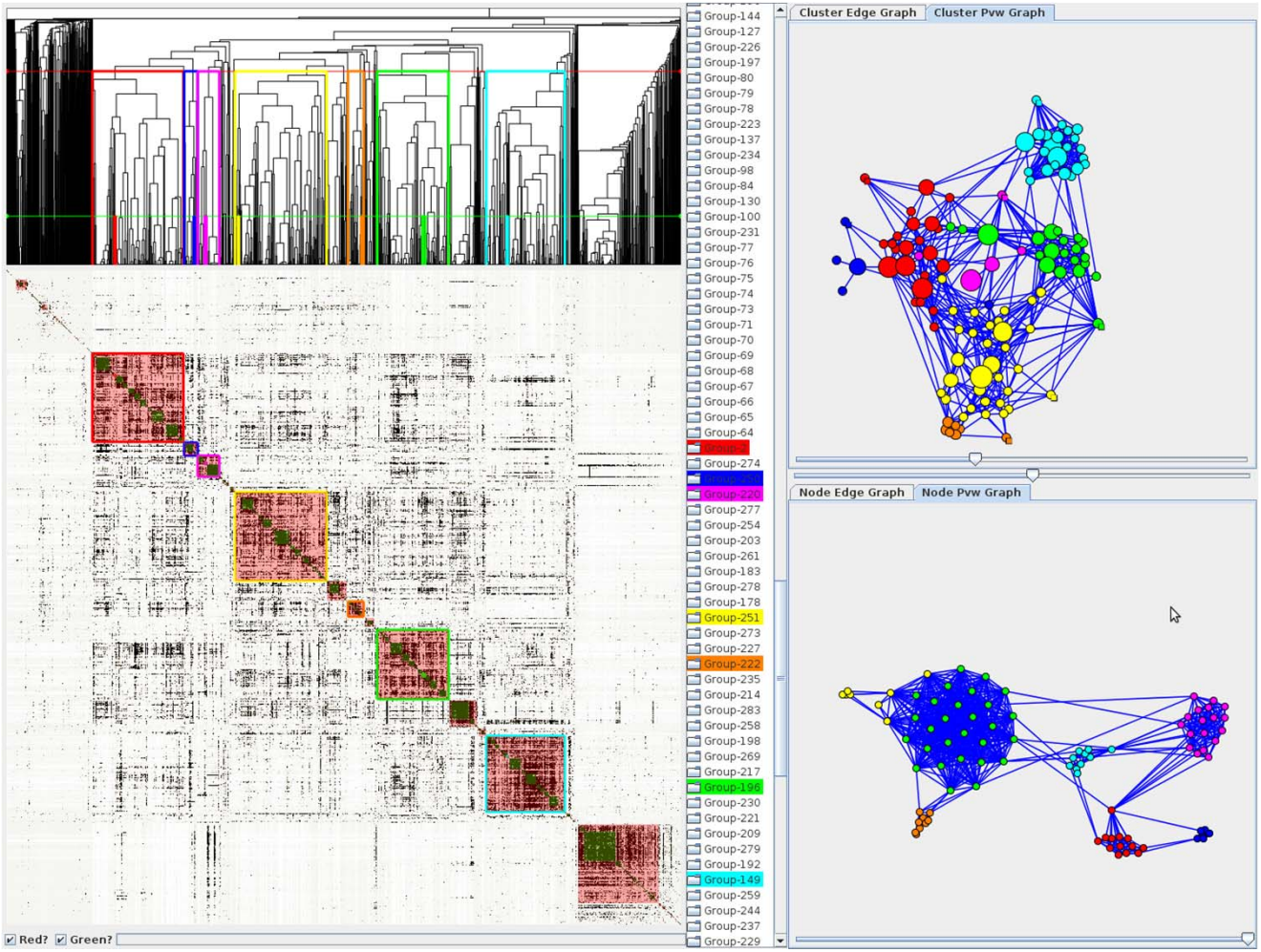
Clicking a particular cluster node in the network, or the corresponding green block in the $p^{\{v,w\}}$ plot will add that cluster's nodes to the bottom panel.

IGNITE Example



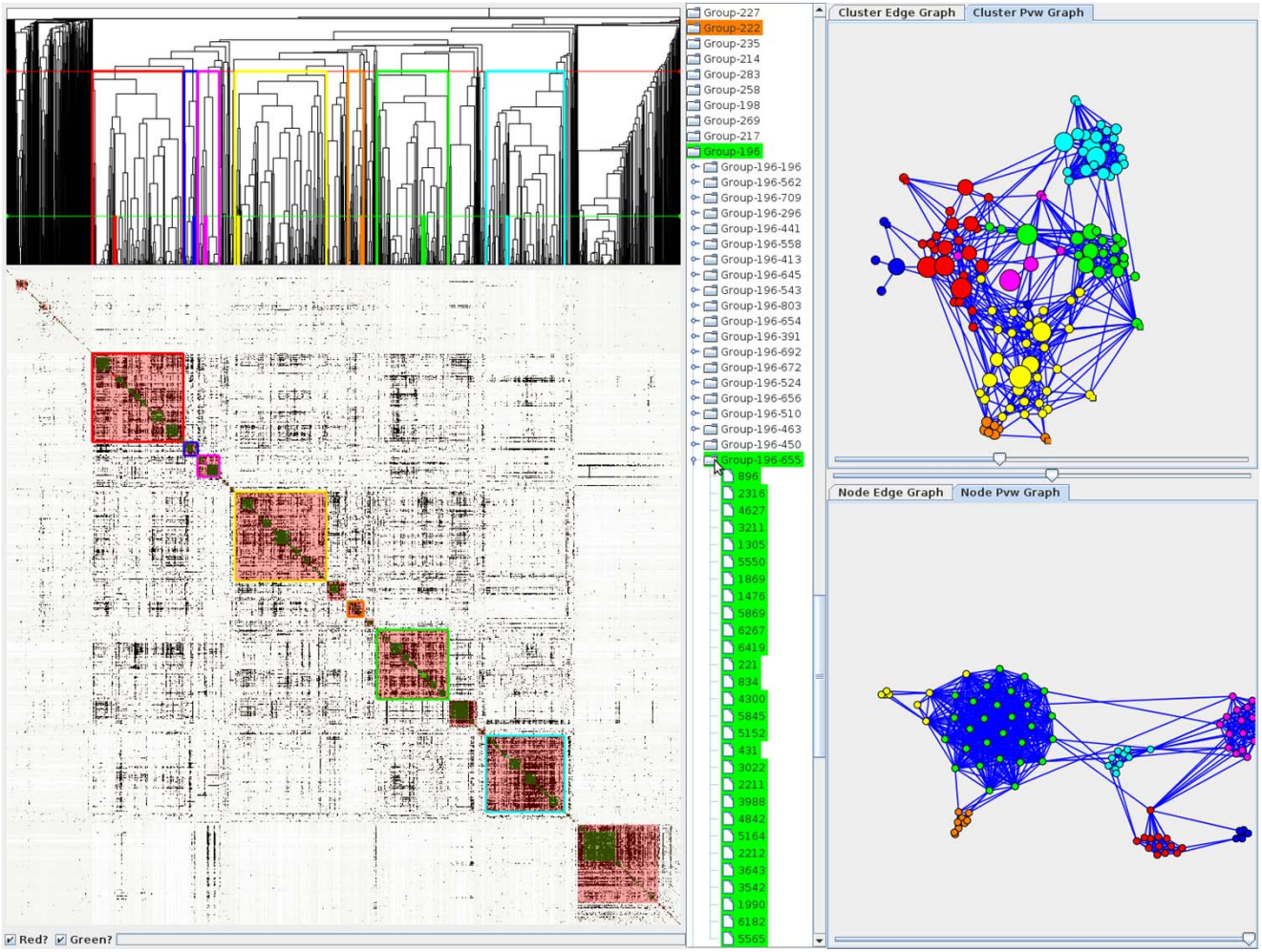
As with the cluster graph, right click to run the layout using the displayed links.

IGNITE Example



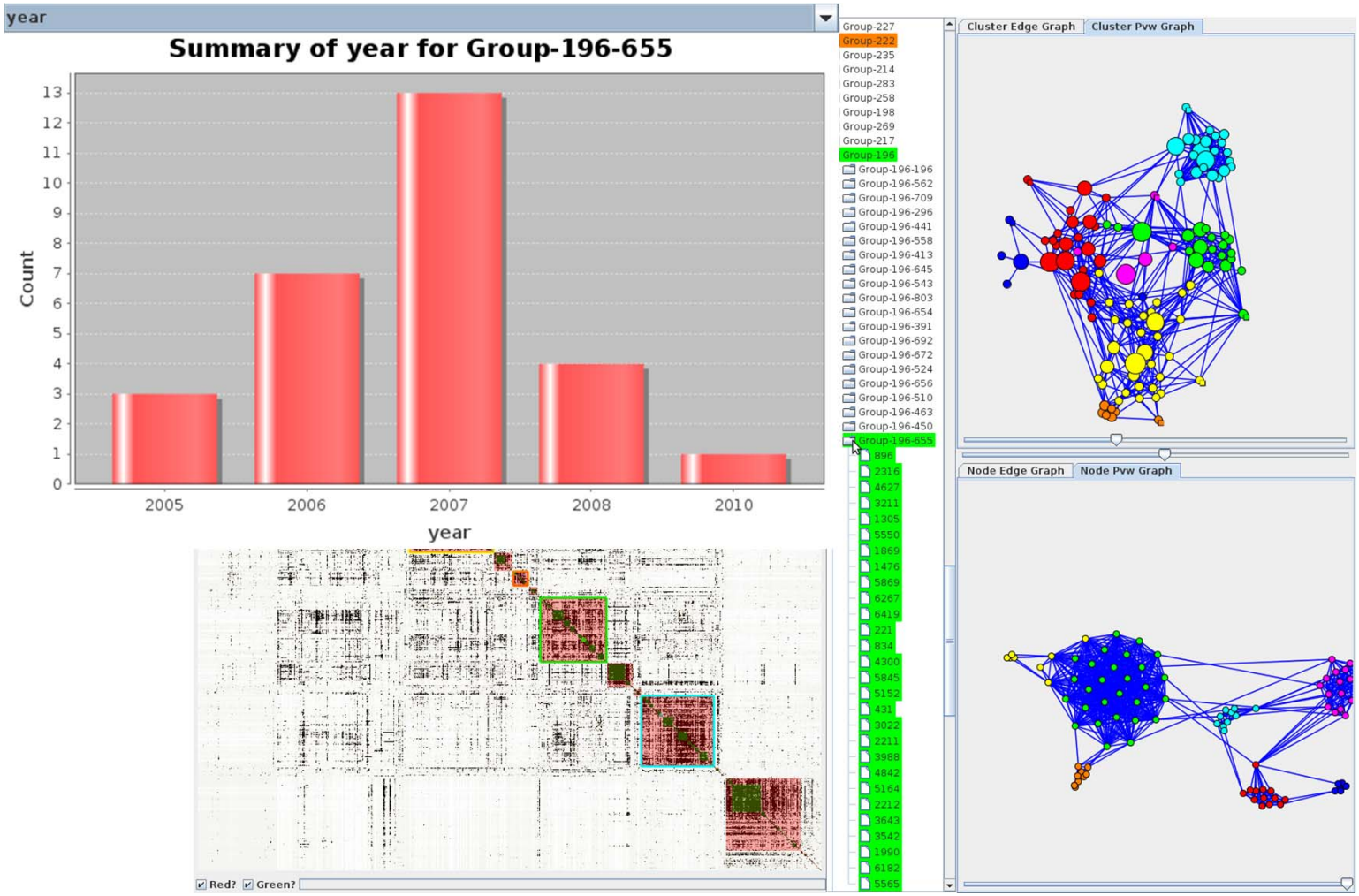
Switching to the Pvw Graph, the Slider is now on a logit scale.

IGNITE Example



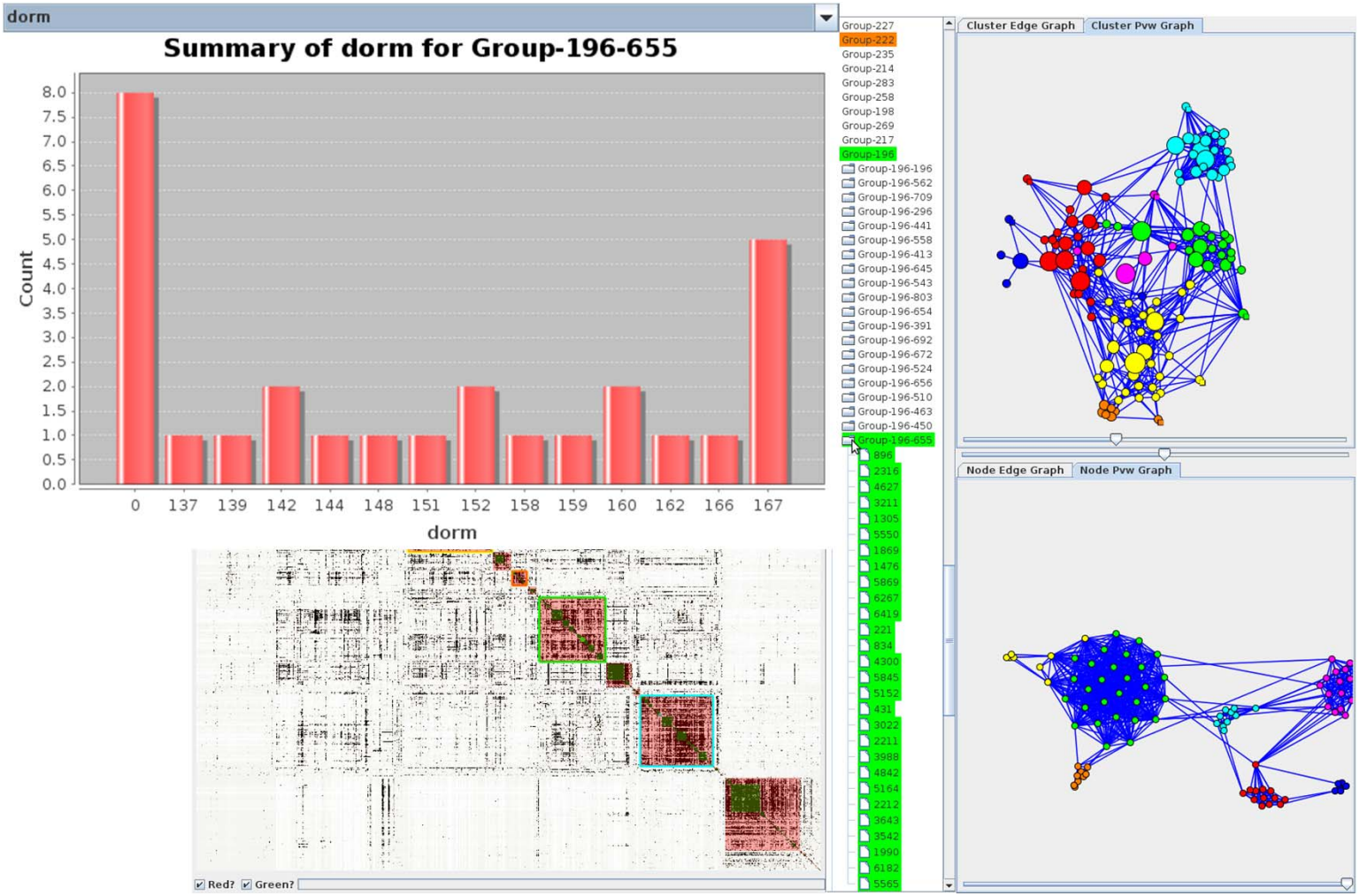
We can expand the tree to explore the group structure, and see which particular nodes fall into which groups. When nodes or groups have names, they're displayed here.

IGNITE Example

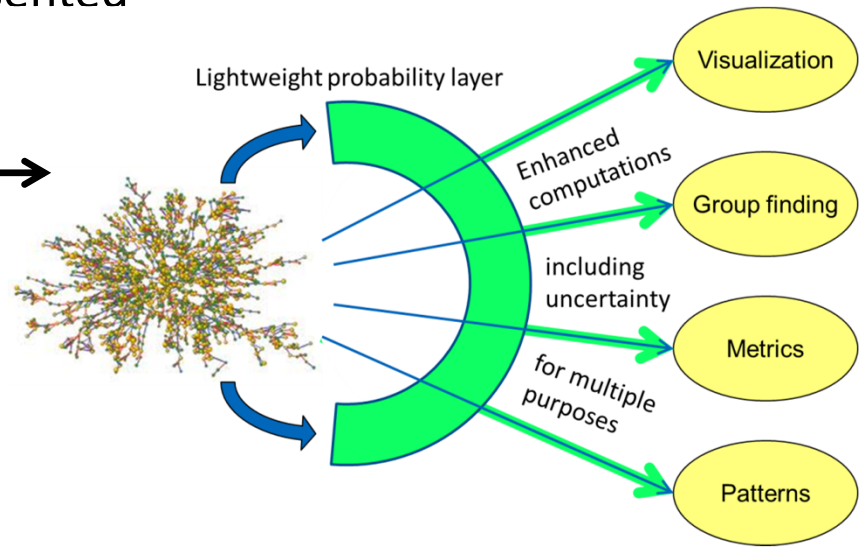


Right click on an entry in the tree for an option to show summary statistics when attributes are available.

IGNITE Example



Selecting different attributes, this group appears to mainly be made up of students in dorm 167 graduating in 2007.

Summary

- Community detection problem
  - Find communities (or groups) of related nodes in networks
  - Extensive literature, many algorithms, but they make hard calls
- Augment hard calls with $p_v(C)$ (probability v is in group C)
- Replace hard calls with $p^{\{v,w\}}$ (probability v and w in same group)
  - Soft alternative to traditional methods
  - Fast, accurate approximation presented
- IGNITE tool based on $p^{\{v,w\}}$
- Basis for further applications →
  - Tracking groups on dynamically evolving networks
  - Distance metric for persistent homology network analysis