



# Detecting of Anomalous Behaviour in Grids

---

Albert Y. Zomaya

*Advanced Networking Research Group*

*School of Information Technology*

*The University of Sydney*

[zomaya@it.usyd.edu.au](mailto:zomaya@it.usyd.edu.au)

<http://www.it.usyd.edu.au/~zomaya>



# Grids

---

- Grid technologies support the sharing and coordinated use of diverse resources in dynamic VOs.
- The creation, from geographically and organizationally distributed components, of virtual computing systems that are sufficiently integrated to deliver desired QoS (Foster, I., Kesselman, C. and Tuecke, S. *The Anatomy of the Grid: Enabling Scalable Virtual Organizations*. *International Journal of High Performance Computing Applications*, 15 (3), pp. 200-222, 2001).



## Grids (2)

---

- Grid **concepts** are critically important for **commercial computing** not primarily as a means of enhancing **capability**.
- A solution to new challenges relating to the construction of **reliable**, **scalable**, and **secure** distributed systems.
- Enterprises must reintegrate (with **QoS**) distributed servers and data resources, addressing issues of navigation, distributed security, and content distribution inside the enterprise, much as on external networks.



# Networks

---

- Contain various physical entities
- Include communicating devices
  - Routers
- And communication links between them
  - Ethernet
  - Wireless
- Can be represented in a more abstract form.



# Graph Representation

---

- Weighted digraphs
- Incorporate
  - Network State
  - Bandwidth/Latency
  - Traffic (bandwidth)
- Time series of graphs, representing sampling at periodic intervals



# Network Management

---

- Adjacent and connected networks affect the local network.
- Knowing the state/behaviour of these networks assists in network management:
  - Intrusion detection
  - Abnormal behaviour
  - Traffic monitoring
- These networks are not under our control!



# External Network

---

- Adjacent/connected networks that we wish to monitor, but cannot.
- Existing methods of monitoring include:
  - Specialised probes
  - Traceroute (<http://www.traceroute.org>)
  - SNMP (<http://www.snmp.com>)
- Issues:
  - Additional network requirements
  - Additional network traffic, that affects the very thing we are attempting to measure.



# Solution

---

- Use *existing* traffic on the monitored network as an indicator of external behaviour.
- Passive approach
  - No additional traffic
  - No additional requirements
- Turns a networking problem into a modelling problem.
- Used in road traffic modelling.



# Topological Change

---

- For large networks, happens relatively slowly
  - Order of days, or weeks, not seconds
  - Relative to the laying of cables etc.
- Can determine the union of possible topologies via an existing method eg. traceroute
- Investigated at a future stage of this research.



# Model

---

- Inputs
  - Packet information:
    - Source, destination addresses
    - TTL
  - Where they impact the monitored network
  - Possible topologies
- Outputs
  - State of each link in the network
  - Bandwidth used on each link in the network
- Run at periodic intervals to provide a changing picture of external events



# Algorithm

---

- Use a genetic algorithm(GA) to model the state of each link in the network.
- Use a shortest path routing algorithm to convert to a set of routing tables
- Use the set of known traffic that arrives correctly as the fitness function.
- Propagate each packet outwards from the arrival and exit points.



# Evolutionary Algorithms

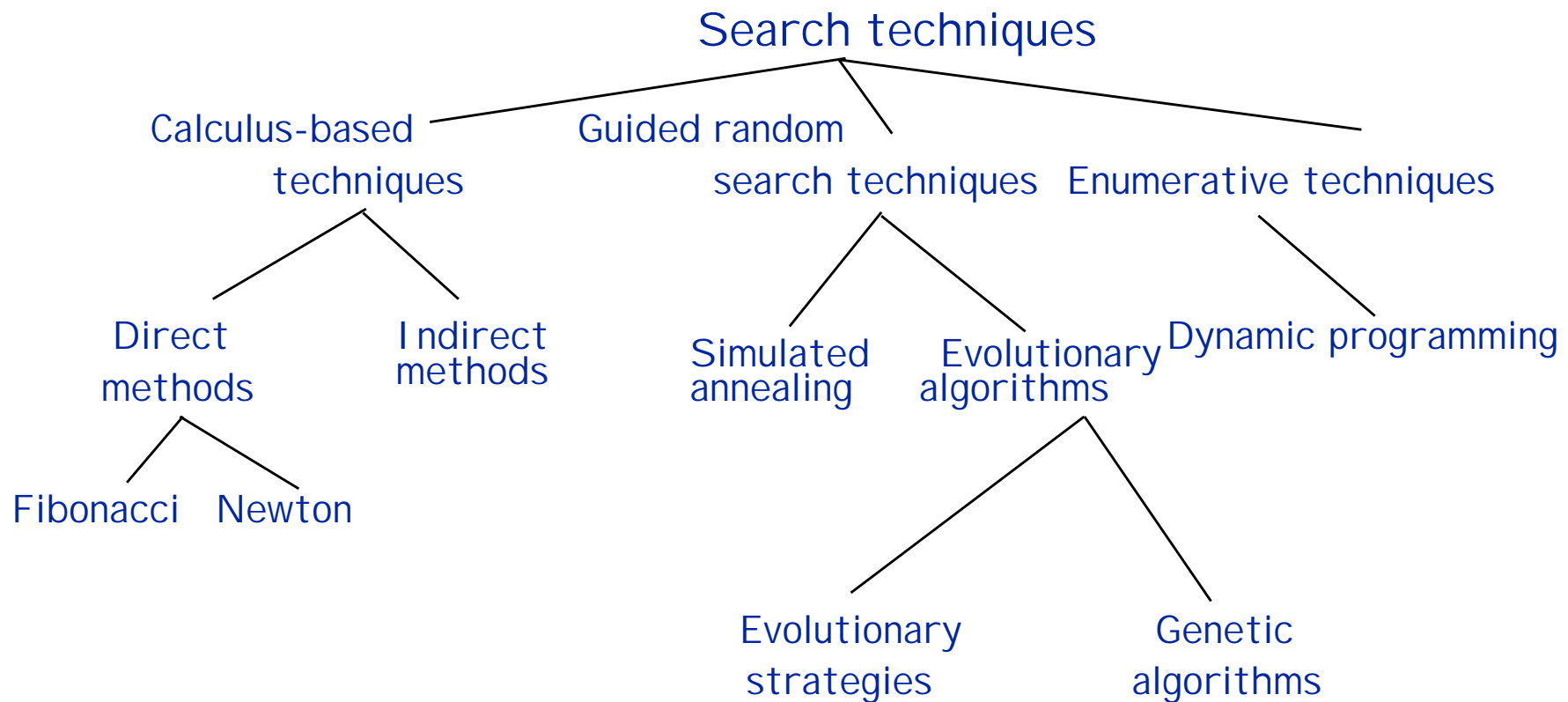
---

- Genetic Algorithms –

- J.H. Holland. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor, MI , 1975.
  - D.E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA, 1989.
- ✓ *Genetic algorithms are more robust, global, and more straightforward to apply.*



# Classes of Search Techniques



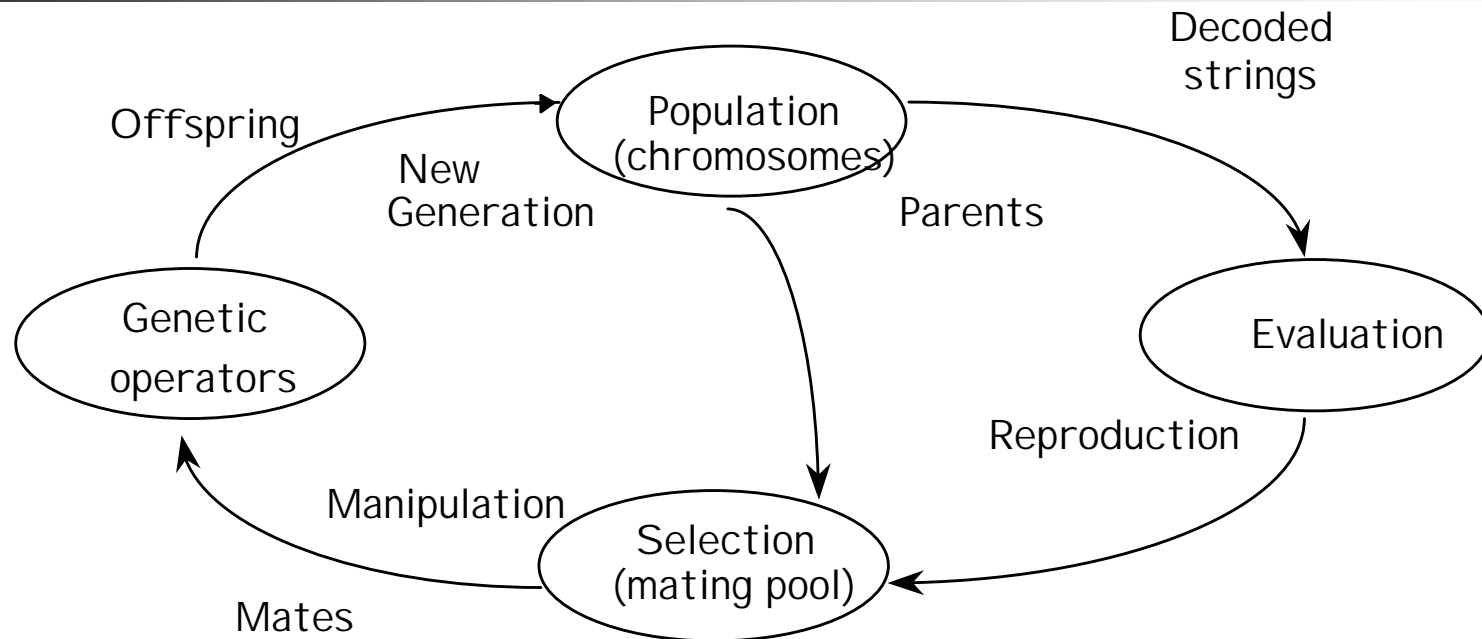


# Auxiliary Information

---

- Many search techniques require **auxiliary** information in order to work properly.
- GAs are blind! Require **payoff** values associated with individual strings.
- By not using auxiliary information a broadly based scheme can be developed.
- The **refusal to use specific knowledge when it does exist can place an upper bound on performance of an algorithm** when compared with methods designed for that problem.

# GAs - Basic Components



- Selection according to fitness is the source of **exploitation**
- The mutation and crossover operators are the sources of **exploration**
- As the mutation rate is increases, mutation becomes more disruptive until the exploitative effects of selection are completely overwhelmed.



# Model

---

- 4 models built, with increasing layers of complexity.
- Each model contained a population of solutions.
- Each solution could be converted into a network representation for the past interval.
- Each solution could be evaluated with respect to the known traffic.



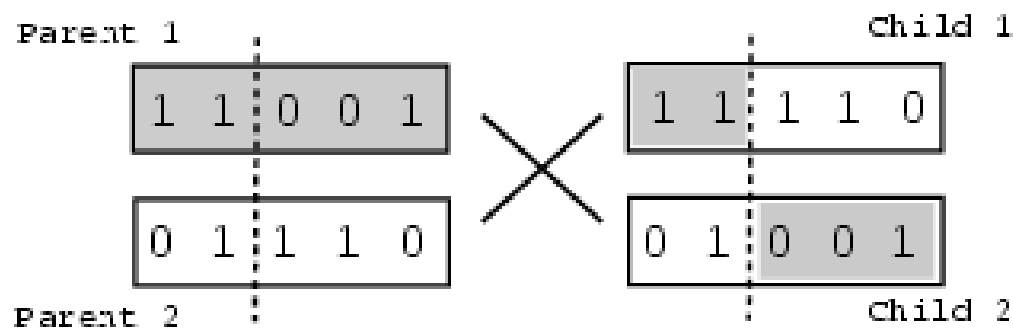
# Chromosomes

---

- Routing chromosome:
  - Represented all the routing tables in the external network.
  - Ignored link state.
  - Array of integer values.
- Topology chromosome:
  - Represented the state of each link in the network.
  - Generated routing tables using Floyd-Marshall algorithm.
  - Array of boolean values.

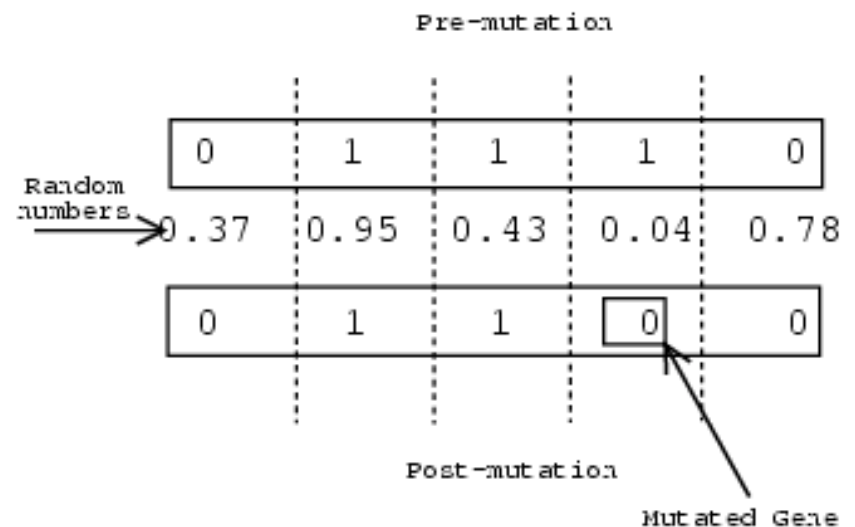
# Crossover

- Ensures traits of both parents can survive.
- Both chromosomes are fixed length
- Standard one-point vector crossover was used.



# Mutation

- Adds diversity to the population, escapes local optima.
- Alter a gene with a given probability (around 5%)
- Change to a new value





# Testing

---

- Use ns-2 to simulate data sets and provide both internal and external network traces.
- Run the model over the internal trace to produce a predictive trace.
- Compare the actual and predicted networks using:
  - Nam (<http://www.isi.edu/nsnam/>)
  - Quantitatively measure accuracy



# Accuracy Measurement

---

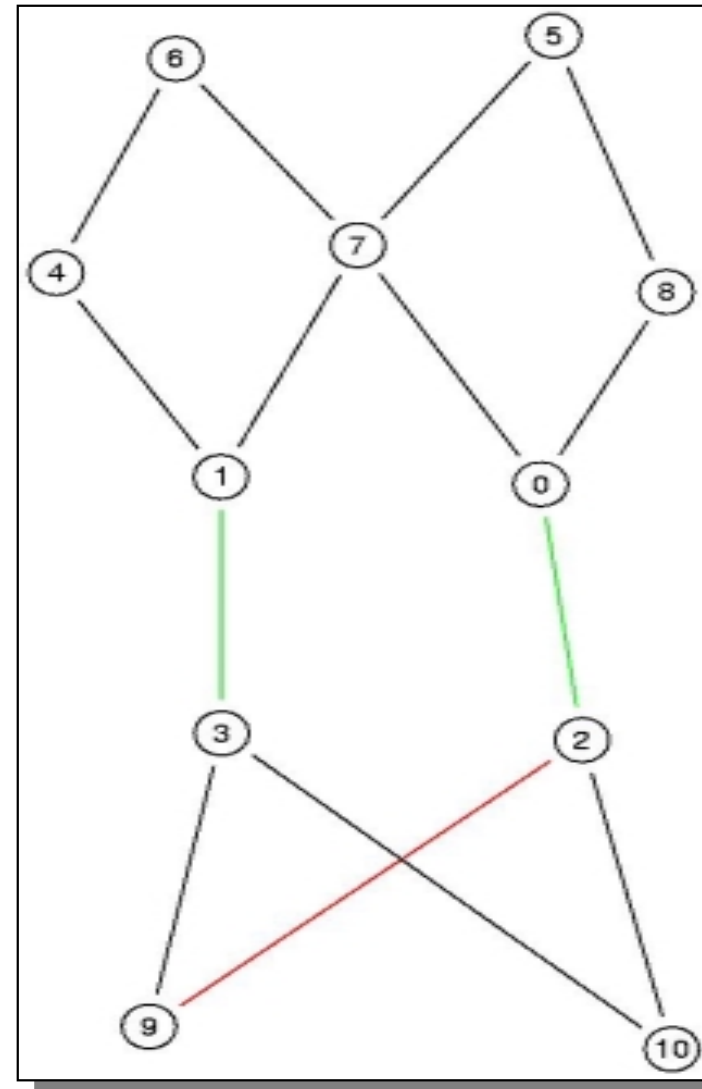
- Averaged over all links in the external network
- Measured at periodic intervals

$$Accuracy_{traffic}(A, P) = 1 - \frac{|B_A - B_P|}{|B_A + B_P|}$$

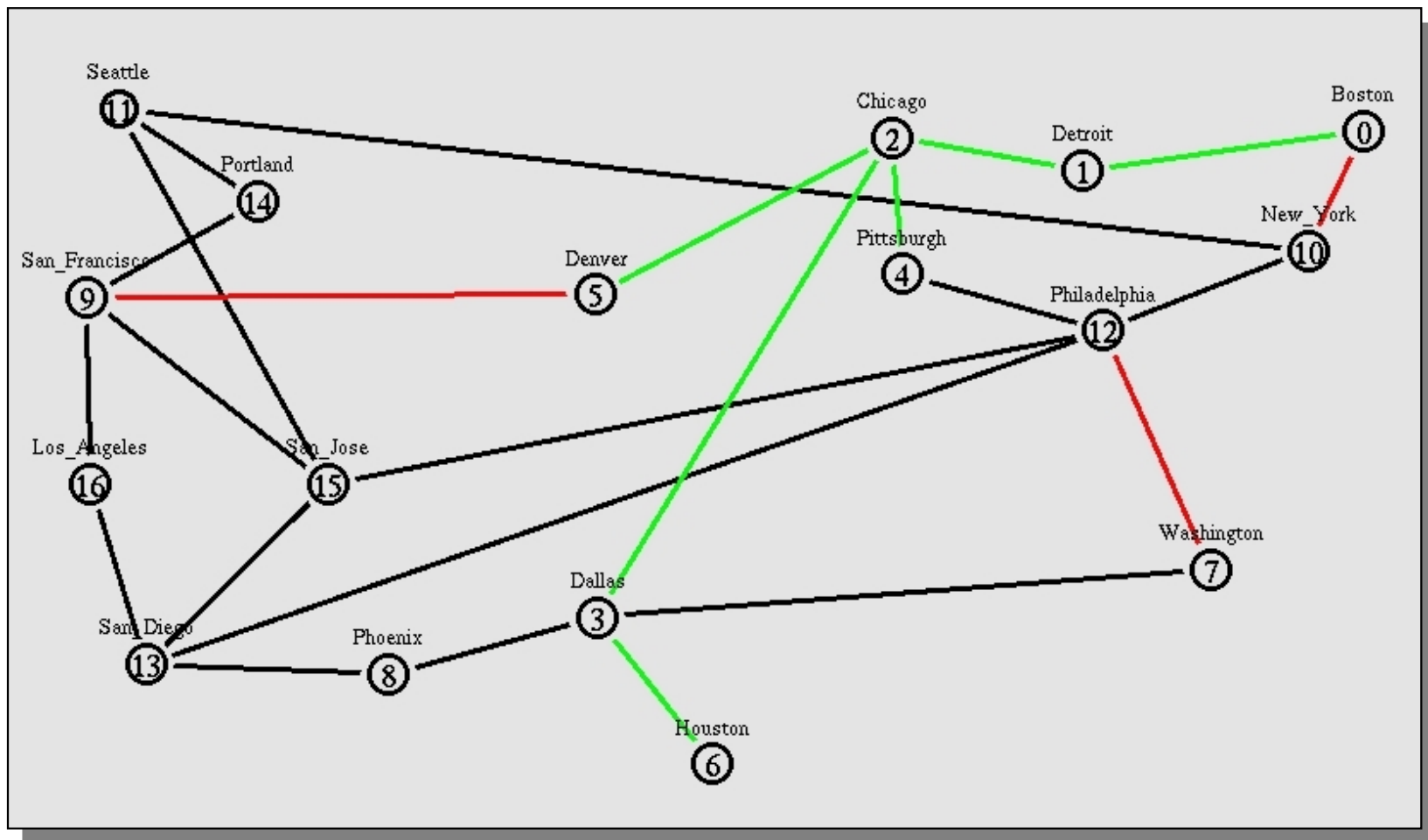
$$Accuracy_{top}(A, P) = S_A \equiv S_P$$

# Case: Standard

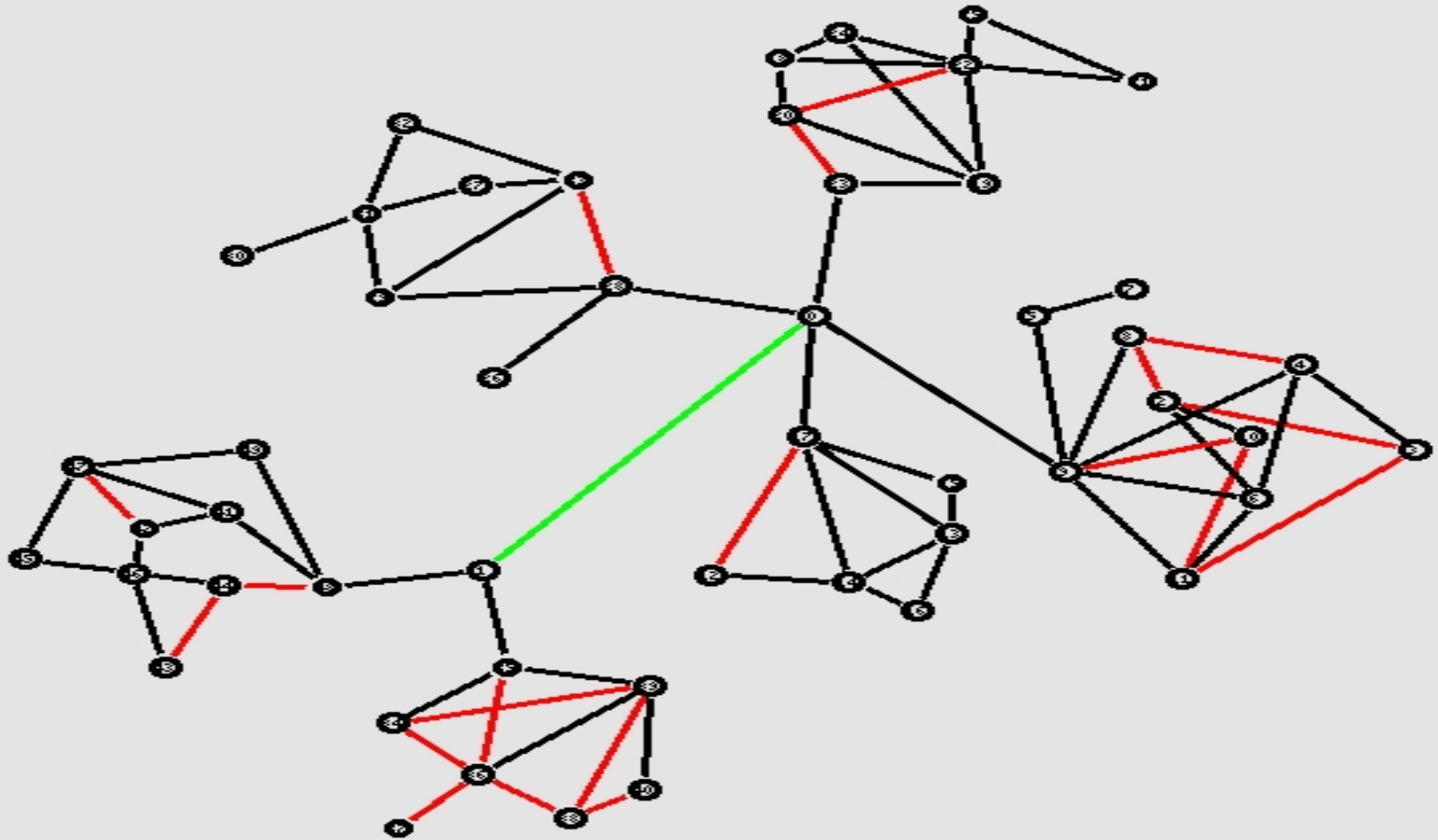
- Monitored links in green.
- Non-existent links in red.
- Valid links in black.



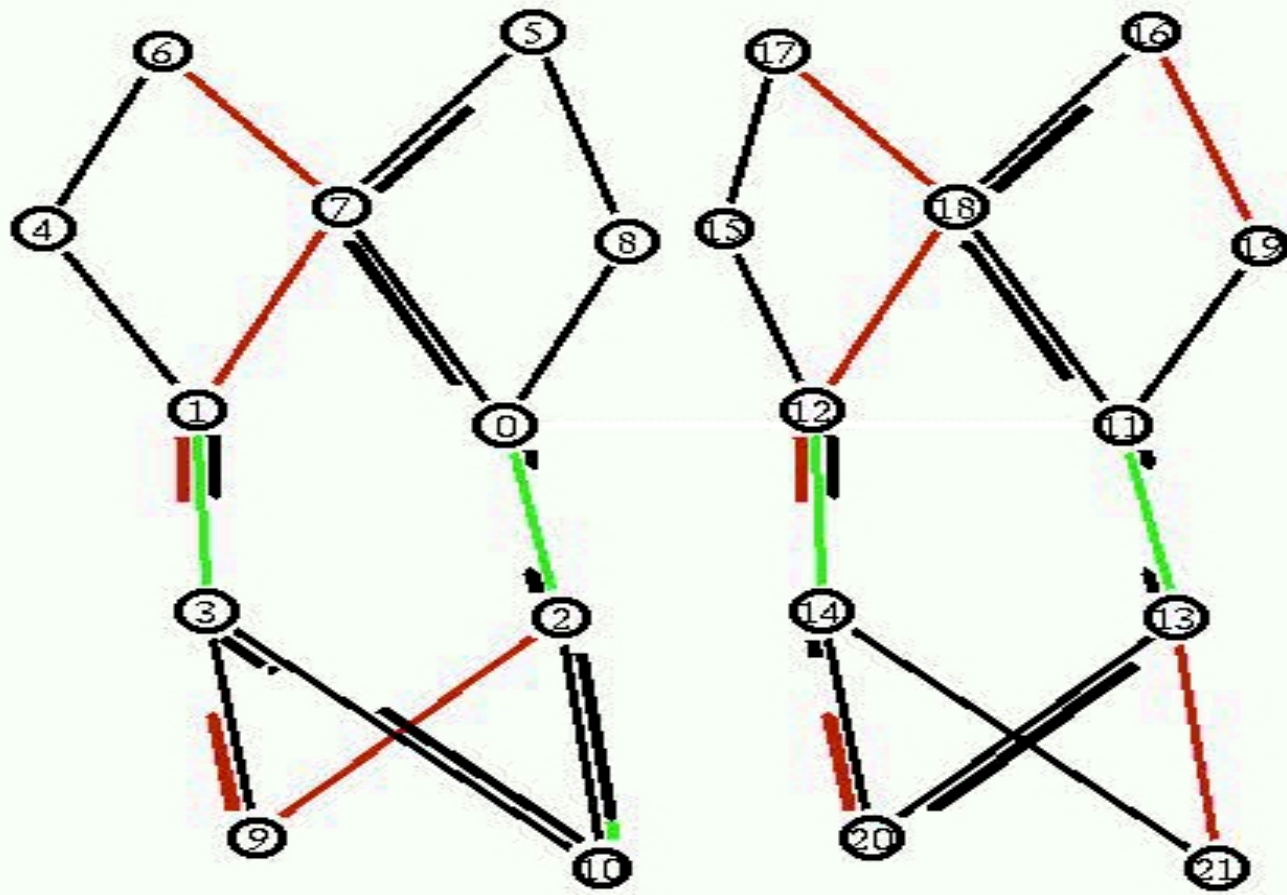
# Case: AT&T Network



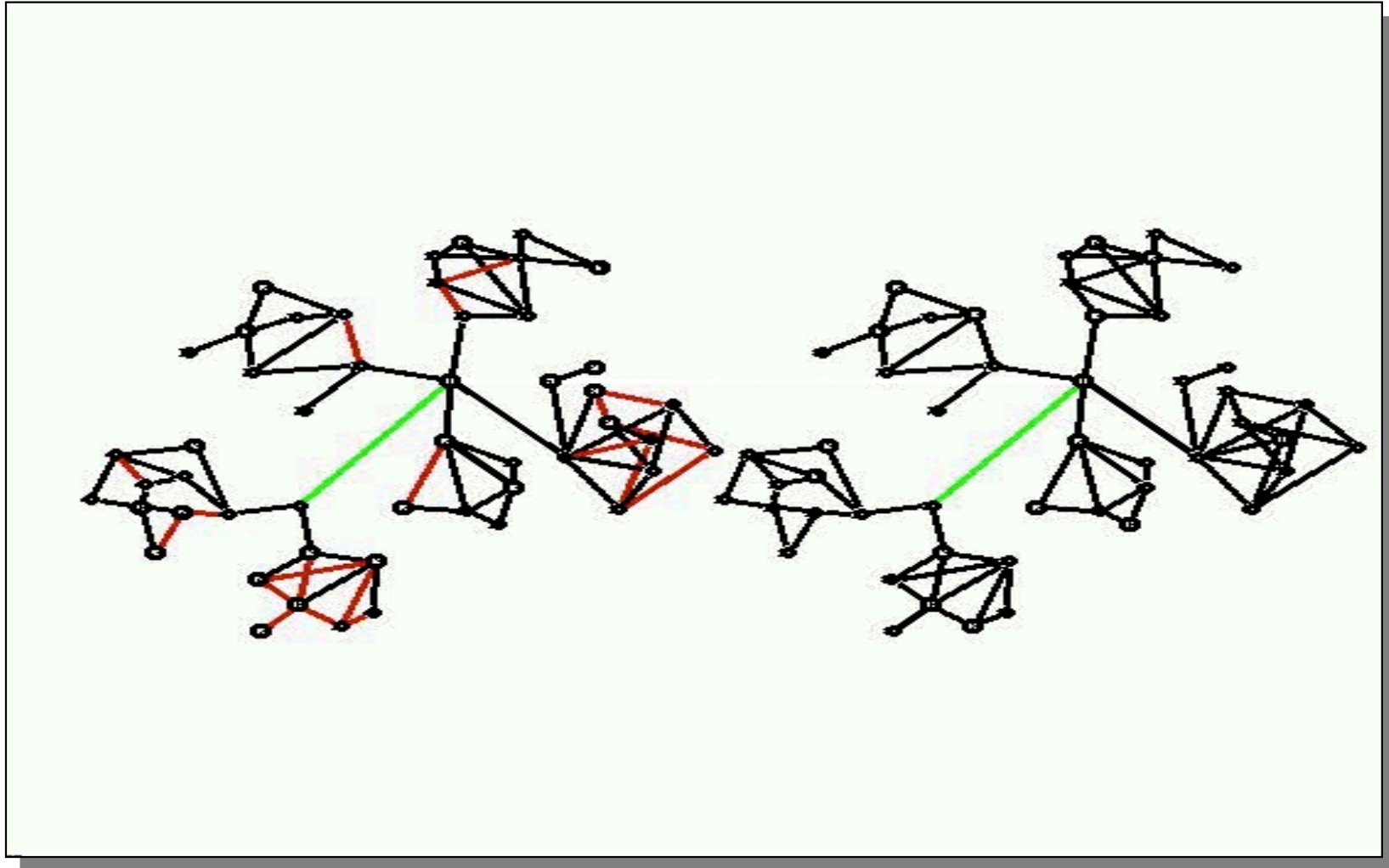
# Case: Large Network



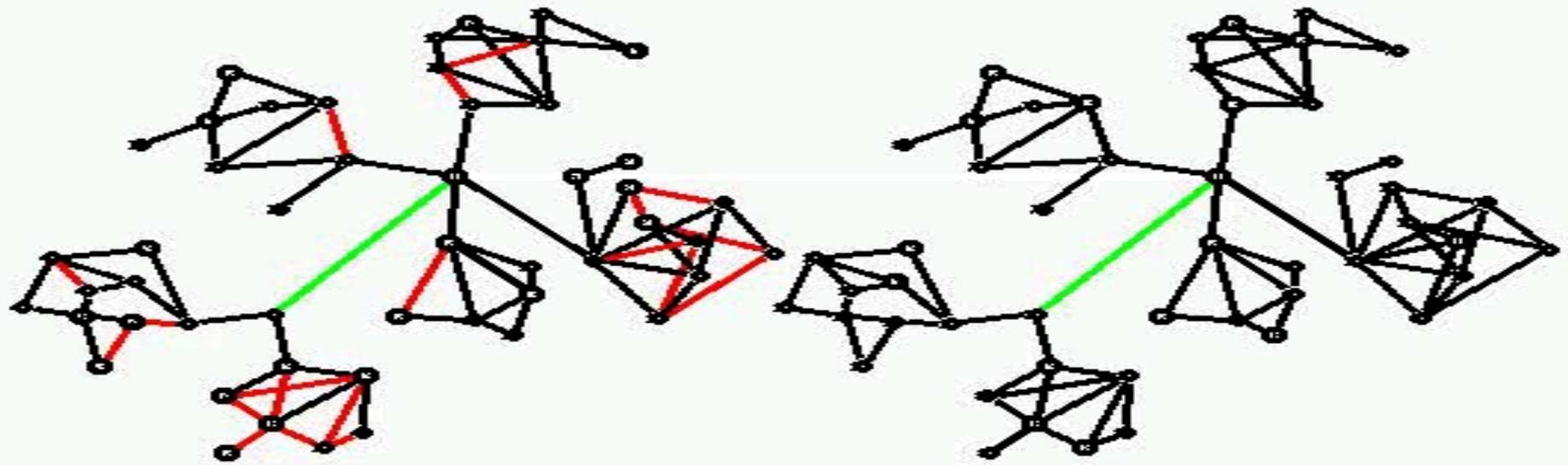
# Standard Visualization



# Large Visualization



# Large Visualization







# Results: Traffic Accuracy

Case	Routing	Topology	Fast	Ttl
Standard	0.684 (0.048)	0.659 (0.041)	0.720 (0.037)	0.744 (0.031)
Large	0.742 (0.028)	0.746 (0.027)	0.780 (0.026)	<b>0.807</b> (0.025)
AT&T	0.558 (0.025)	0.571 (0.018)	0.566 (0.024)	0.621 (0.029)

Format: mean(95% confidence interval)



# Results: Topology Accuracy

Case	Topology	Fast	Ttl
Standard	0.602 (0.042)	0.616 (0.044)	0.606 (0.053)
Large	0.543 (0.024)	0.572 (0.016)	0.576 (0.008)
AT&T	0.648 (0.017)	<b>0.675</b> (0.018)	0.665 (0.027)

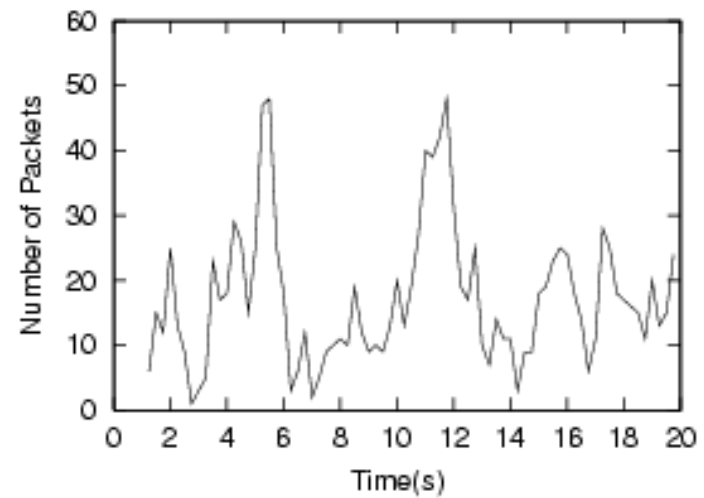
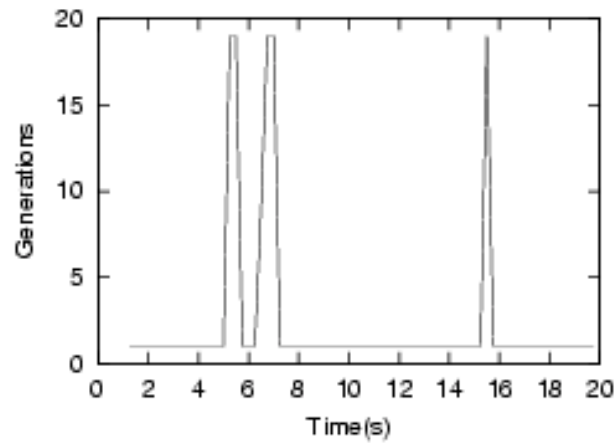
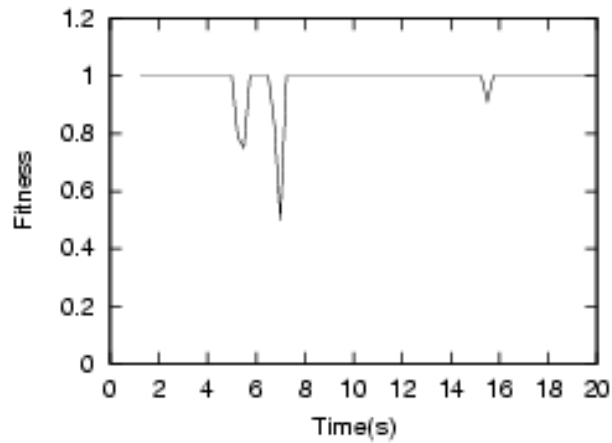
**NB: Routing model didn't predict topology**



# Results: Internal Fitness

Case	Routing	Topology	Fast	Ttl
Standard	0.973 (0.018)	0.972 (0.019)	<b>0.983</b> (0.016)	0.973 (0.027)
Large	0.940 (0.030)	0.990 (0.006)	0.965 (0.021)	0.871 (0.052)
AT&T	0.553 (0.040)	0.609 (0.014)	0.634 (0.017)	0.618 (0.021)

# Internal Fitness: Graph





# Conclusions

---

- Accuracy
  - Topology: Proportional to the relative size of the internal network.
  - Traffic: Proportional to the amount of monitored traffic / centrality of the network.
- Internal fitness
  - Could assist in determining network change.
  - Provides a measure of prediction confidence.



# Future Work

---

- Test on a real network.
- Compare traffic loads with existing techniques such as traceroute, probes etc.
- Model each subnetwork separately using an island model, with a higher layer combining predictions.
- Combine with active techniques to find an optimal trade-off between network cost and accuracy.