Principles of Scalable Dynamic Visual Analytics

H. V. Jagadish, Comp. Sc.

and

George Michailidis, Stat.

Univ. of Michigan, Ann Arbor

Broad Research Goal

Develop fundamental principles

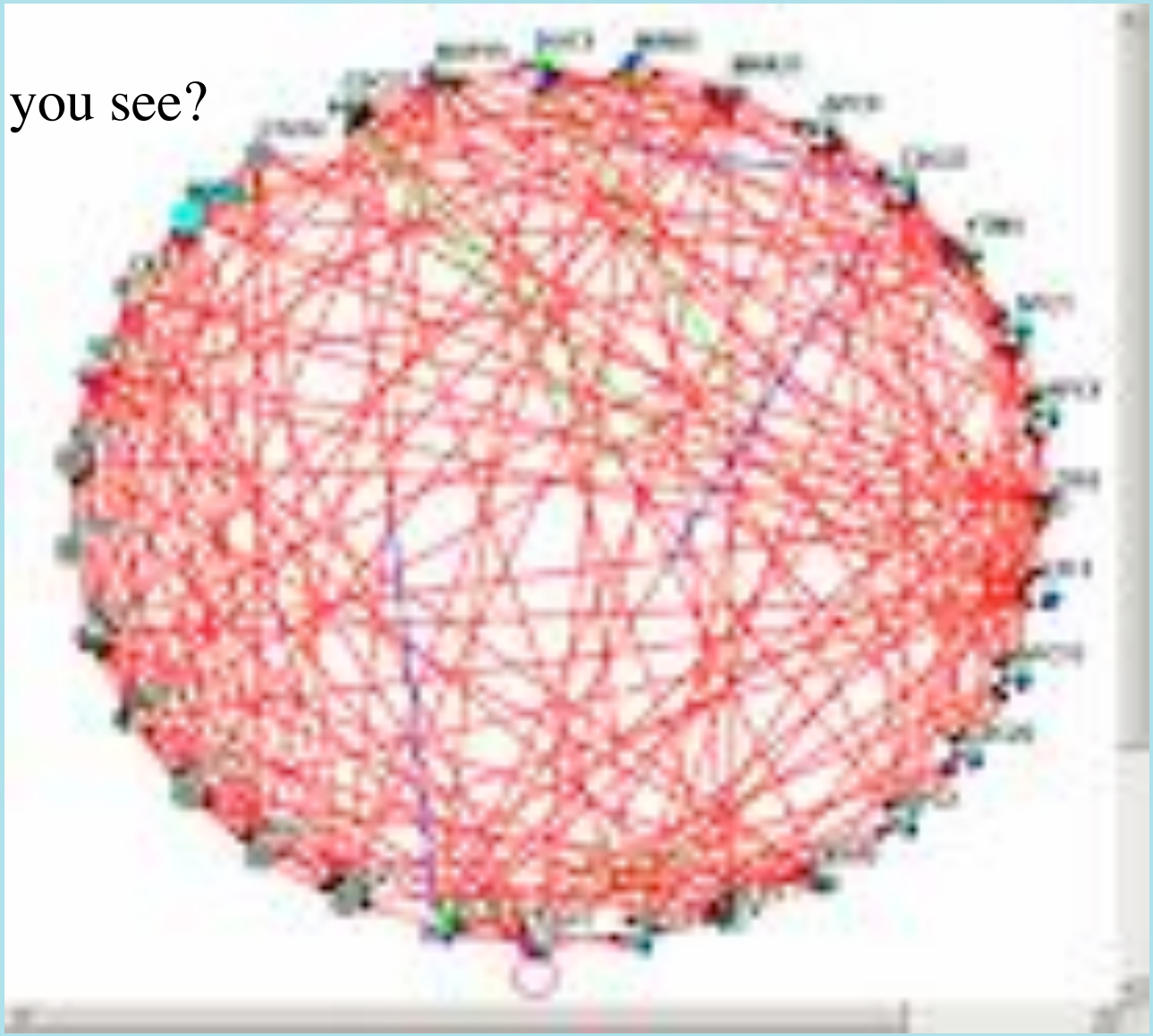
and an effective tool set

for simplification and presentation

of large complex dynamic data sets

in a manner that facilitates visual analysis by humans.

What patterns do you see?



Focused Research Agenda

- Limit attention to graph data.
- Permit attributes on nodes and on edges.
- Selection and aggregation are two natural operators to reduce large graphs.
  - But, based on what?
- Are there other useful operators?
- How do we deal with change?

Problem Definition

- Let $G_t=(V_t,E_t)$ denote a graph at time t.
- Associated with the nodes & edges are numerical and categorical features (X_t,Z_t).
- For example, time course protein expression data associated with a protein interaction map, or gene expression data associated with a signaling pathway.

Challenges for Visual Analytics

- Size of the graph.
- Dimensionality of feature sets.
- Multiple time points.

Useful Operators for Addressing these Challenges

- Aggregation -> reduces the size of the graph, allows to summarize information over time, etc.
- Selection -> feature based (e.g. select graph elements that satisfy a specific predicate) or graph based (e.g. select high degree nodes).
- Manipulation -> adds information for analytics tasks (e.g. classification of nodes).

Towards an Operator Language: some Challenges

- Issue: for many visual analytics tasks a sequence of operators is required (e.g. selection of highly expressed proteins, pathway build, network alignment for comparisons across experiments).
- Goal: study properties of operators and their classes (aggregation/selection/manipulation) and the composition of members of the class and across classes.

A more complex class of operators: feature scaling

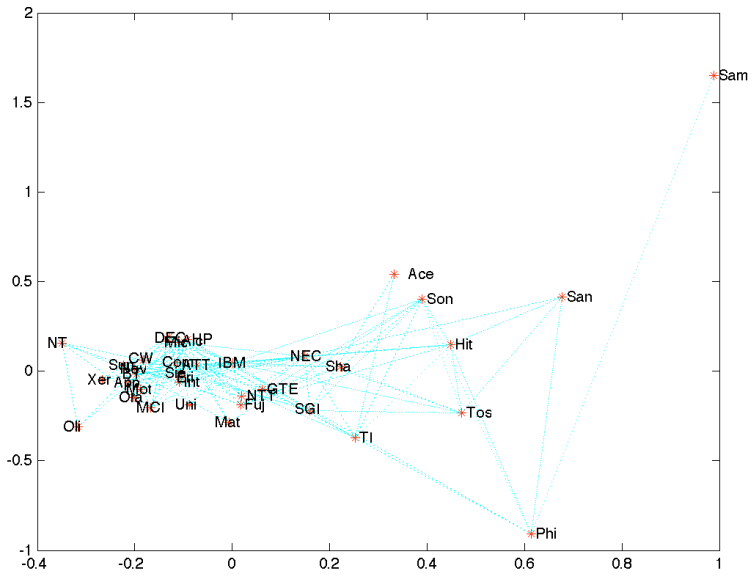
- Feature ranking: aim to simultaneously rank binary/categorical/numerical features and graph nodes.
  - Result: the eigenvector corresponding to the second smallest eigenvalue of an appropriate constructed Laplacian matrix produces an optimal ranking in the case of binary features and an approximate one for other types of features.

A more complex class of operators: feature scaling (ctd)

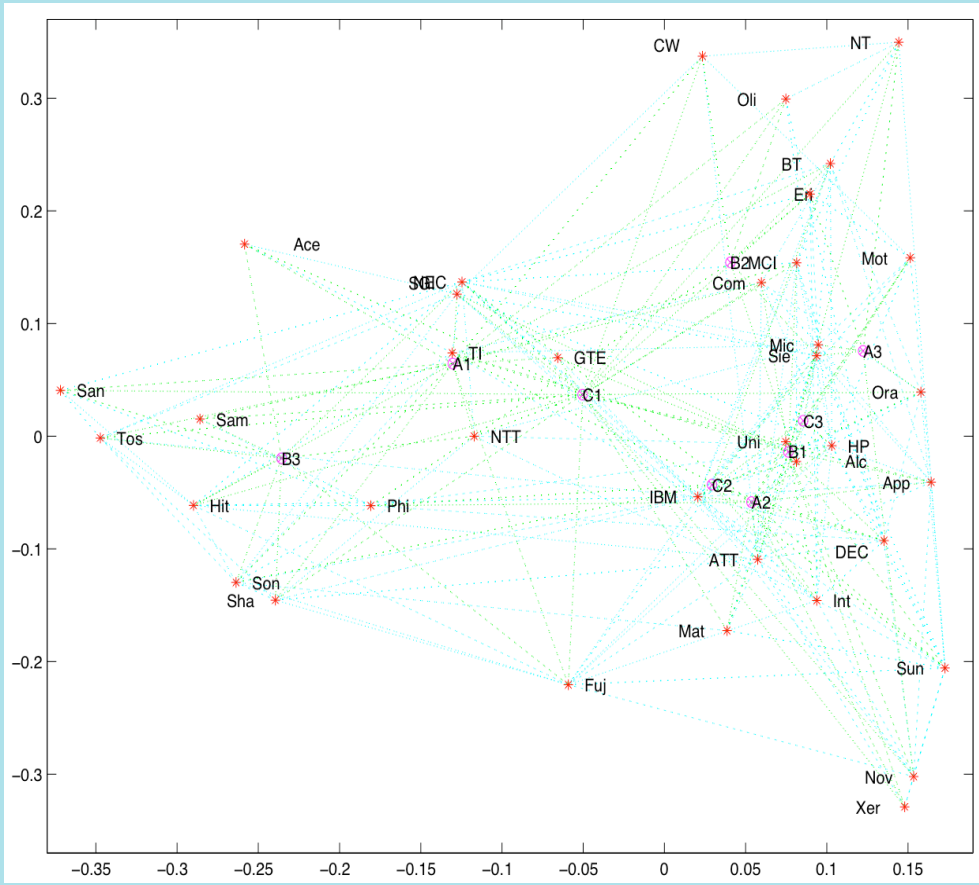
- Dimension reduction: aim to summarize appropriately a large number of features.
  - Eigenvectors corresponding to largest eigenvalues of covariance matrices (numerical features) or collections of contingency tables (categorical ones) provide optimal linear summaries. Extensions to non-linear summaries.

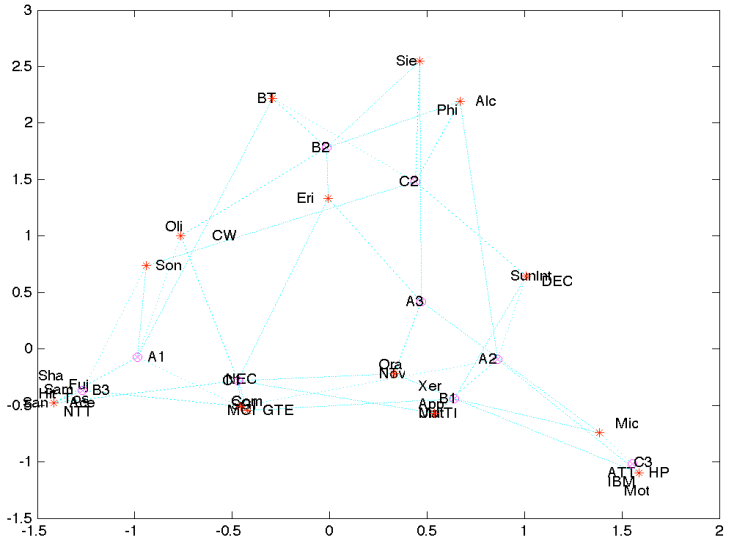
Extensions to fusion operators

- Network structure (G) provides one view (V_1).
- Features (X, Z) provide a different view (V_2).
- Goal: Develop mathematical framework for fusing views ($f(V_1,V_2)$).
- What are the properties of f?
- How to measure the quality of the composite view? What loss functions to use?
- What is the appropriate sequence of operations?



<- Network view





<- Feature based view

Composite view

Operators over time

- Temporal structure (e.g. time course data).
- Goal: extend operator language to capture dynamics.
- Simplest possible operation: averaging over time. Can we do better?
- In many instances, exponential weighted moving averaging proves useful. Issues: how to choose the weights?
- Temporal selection in place of temporal aggregation?
- Reduction to (small) “k” rather than reduction to 1.

Application and Outreach

- Both senior investigators have vibrant ongoing collaborations with biologists.
- “hairball” protein interaction network is an ideal target problem.
- We will demonstrate use of visual analytics for protein data using FODAVA techniques.

- Plus, of course, other applications through FODAVA community.