

KDD 2013 Workshop on
Interactive Data Exploration and Analytics (IDEA)

A Process-Centric Data Mining and Visual Analytic Tool for Exploring Complex Social Networks

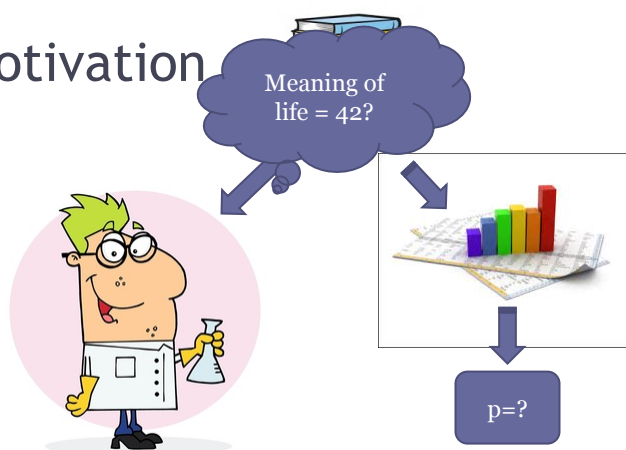
Denis Dimitrov, *Georgetown University*,
Prof. Lisa Singh, *Georgetown University*,
Dept. of Computer Science

Prof. Janet Mann, *Georgetown University*,
Dept. of Biology



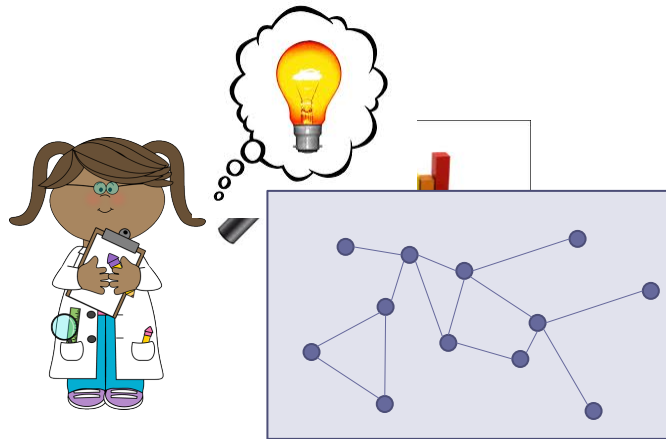
GEORGETOWN UNIVERSITY

Motivation



Analysis driven by subject matter theory and expert
knowledge of domain

Alternative approach



Data driven analysis

More specifically

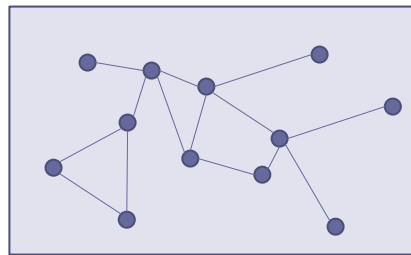
- Develop a visual mining tool for analyzing graphs.
- Observational scientists study graphs to better understand sociality, community structure and changing dynamics subjects.



Existing Graph Visual Analytic Tools

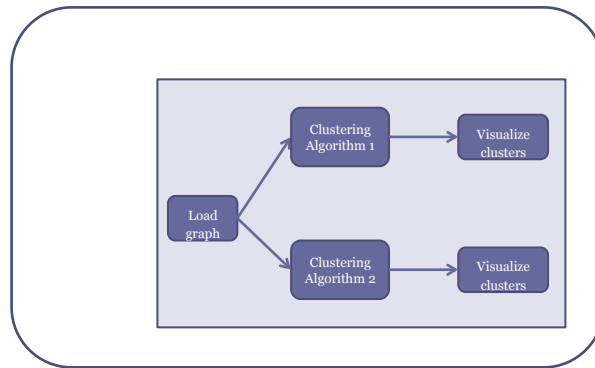
- Jung: a toolkit containing a number of data mining algorithms
- Prefuse, Gephi have more extensive visualization support
- Guess has a query language for manipulating graph visualization

Data-centric tools



- Built around the data, with which the user interacts directly
- The notion of analytical process is not explicitly present

Process-centric tools



- Built around the analytical process
- Users interact with the data by defining, configuring, and running the process

Workflow tools

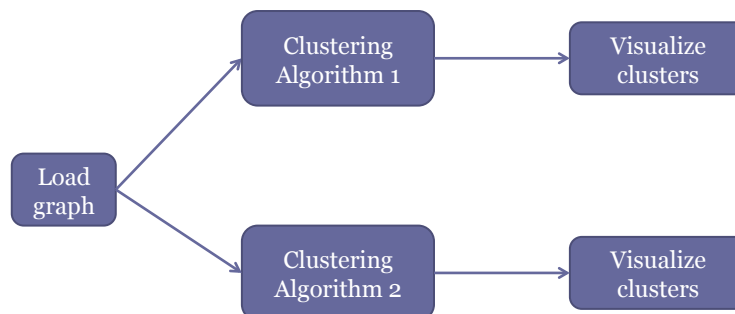
- Existing tools:
 - Orange
 - RapidMiner
 - Weka
 - Knime
 - Amira and Voreen (volumetric data visualization)
 - Viztrail (provenance)
- Our goal: integrating workflow concepts with graph analysis tasks

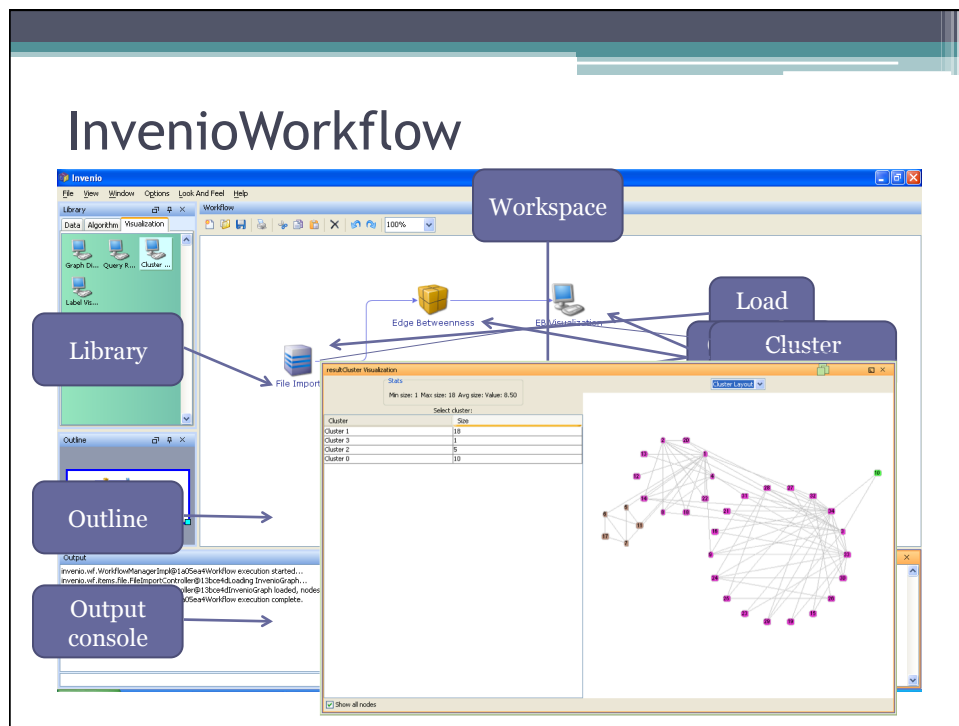
Benefits

- For scientists without extensive programming knowledge:
 - Ease of use (visual vs. programmatic)
 - Flexibility in defining the analytic process by combining desired individual tasks as building blocks
 - Experimenting with different algorithms
 - Building the analytic process incrementally and iteratively



Workflow for a sample task



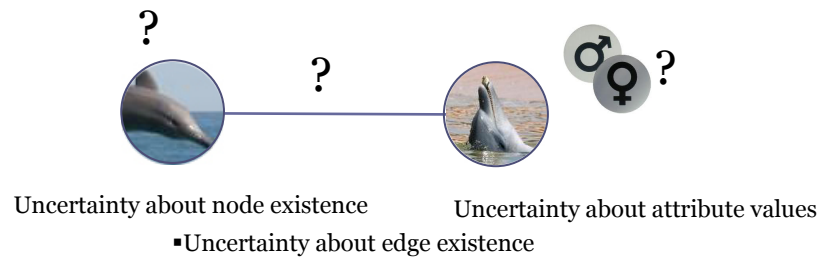


Use cases - dolphin social network

- **Dataset:**
 - Dolphin social network identified during approx. 30 years of study of a dolphin population in Shark Bay, Australia
 - 800 dolphins (nodes) and 29,000 social interactions (edges)
- **Task:**
 - Identify new friends and repeat friends of a specific dolphin between the years 2010 and 2009 by querying for the difference and intersection of the dolphin's ego-network

Extended graph data model

- Graph + Uncertainty = Uncertain Graph :



Use cases - CORA citation dataset

- Dataset:
 - 2708 machine learning papers (nodes)
 - 5429 citations (edges)
 - 7 possible paper topics (label attribute)
- Task:
 - Compare the results of 2 different node labeling algorithms, which predict the topic of each publication

Conclusion: contributions

- A prototype process-centric, visual analytic tool to aid scientists in data-driven analysis
- A workflow process that includes visual, data mining, and graph query widgets for custom, exploratory analysis of network data
- Integration of a graph query engine into the workflow process
- A demonstration of the utility of the proposed workflow design using a complex dolphin observation network and a citation network

Future directions

- Increase the number of data mining and other exploratory tasks supported by the tool
- Optimize performance for graphs that do not fit into main memory
- Develop more sophisticated widgets related to time-evolving networks and information diffusion
- Perform usability study

Questions



Denis Dimitrov: dd322@georgetown.edu

Lisa Singh: singh@cs.georgetown.edu

Supported by:

NSF - Grant Numbers: 0941487 and 0937070

ONR - Grant Number: 10230702