Interpreting Card-Sorting Data with Categorization Graphs

Steven F. Wolf, Michigan State University
D. P. Dougherty, Michigan State University
Gerd Kortemeyer, Michigan State University

– Abstract –

On its 30th anniversary, we re-examine the seminal paper by Chi et. al. (1), which firmly established the notion that novices categorize introductory physics problems by "surface features" (e.g. "incline," "pendulum," or "projectile motion"), while experts use "deep structure" (e.g. "energy conservation" or "Newton's second law"). The paper has been cited over 3000 times in scholarly articles across a wide range of disciplines. Yet, a clear relation of the statistical underpinning for Chi et. al.'s original categorization experiment have remained elusive. We propose here a statistical model of the categorization cognitive process. Application of the method to an expanded Physics education categorization data set gives fresh insight into the cognitive structures of physics experts and novices.

– Motivation –

- Nobody has straightforwardly replicated Chi et.al.'s (1) study
  - Chi's questions are lost
  - Chi's exact analysis method is lost
- Conclusion bears examining considering above difficulty
- Similar (2–6) have used different methods

–Our Study Design–

Build a statistical model in order to compare different categorizations and analyze the groups of problems created by the reviewers

- Model random categorization data
- Use standard graphs
  - Erdös-Renyi–Uniform–Graphs
  - Barabasi–Small World–Graphs
- Use benchmark statistics from Graph Theory to compare models to data
  - Calculate 3-cycle statistics for the data distribution and each model distribution respectively
  - Compare these distributions using KS-Test.

Statistical Analysis

Emperical Cumulative Distribution Function (CDF) for any distribution $D(x)$

$$CDF(x) = \int_{-\infty}^{x} dx' D(x')$$

Kolmogorov-Smirnov Test (KS-Test) compares two CDFs

$$\text{KS-statistic} = \max |CDF[1] - CDF[2]|$$

3-cycles: A sub-graph consisting of three vertices all of which are connected by edges.

Why 3-cycles?

- Describe degree of connectivity of each sample
- Optimize the KS-Test statistic to determine best fitting statistical model

–Categorizations as Graphs–

Properties of categorizations as graphs:

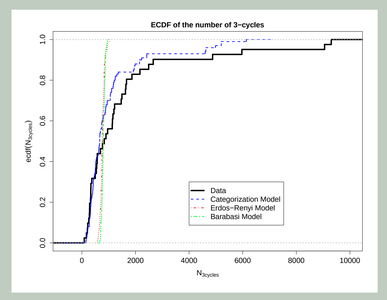
- Problems are vertices
- Edges connect problems in a group
- Graphs are *undirected*
- Calculate the number of 3-cycles for each graph
- Compute the emperical CDF of the number 3-cycles

Toy example: Categorize the numbers from 1 through 10.

Categories for graph to left:

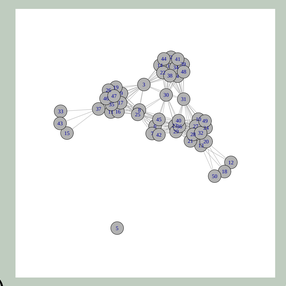
- Even numbers
- Odd numbers
- Prime numbers
- Perfect squares

–Statistical Models and 3-cycles–

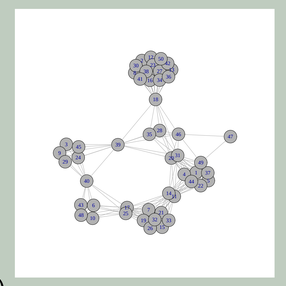


Example graphs from the data distribution and each model distribution

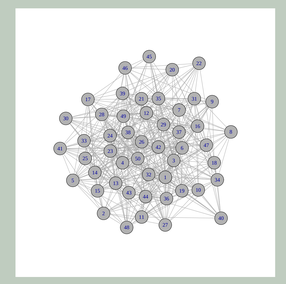
(a)



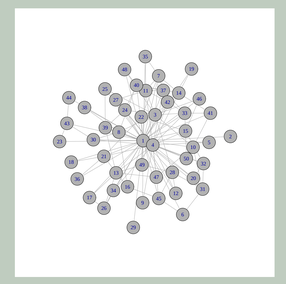
(b)



(c)



(d)



(a) Expert #7's graph
(b) A categorization model graph using optimized parameters
(c) An Erdös-Renyi (uniform) model graph for optimized parameters
(d) A Barabasi (small-world) model graph

– Model Assumptions–

Chi et. al. says:	We say:
Look at category names	Look at problem groupings
Categorization is a deterministic process	Categorization is a random process
Variation points to underlying method	Variation points to random behavior
Use population statistics	**Use sample statistics**

–Categorization Cognitive Model Pseudocode–

```
for each graph
    Q = number of questions
    C = random deviation from binomial distribution
# Create T matrix rows are questions columns are categories
    Initialize T
    X = randomize question numbers
    Y = shuffle list of numbers from 1 to C
# Rule #1:  Every category must be used
    for all j in 1 to C
        T(X(j), Y(j)) = 1
# Rule #2:  All questions must be categorized at least once
    Z = sample the list from 1 to C with replacement Q-C times
    for all j in 1 to (Q-C)
        T(X(C+j), Z(j)) = 1
# Rule #3:  Each question may be categorized more than once
    for all zero elements left in the T matrix
        if (random number from 0 to 1 < beta) T = 1
    # Convert T matrix into adjacency matrix (adj) where
        if T(i,) dot T(j,) > 0
            adj(i,j) = 1
        else
            adj(i,j) = 0
```

–Future Plans–

- Look at other Graph Theory statistics and properties to compare data and models.
  - All pairs shortest path
  - Diameter
  - Transitive closure properties

–Conclusions–

- Based on 3-cycle data we find no significant difference between experts and novices
- Our new categorization model is predictive, summarizes the data, and embodies the "rules" of the cognitive process.
- The number of 3-cycles measures relative clustering vs. hierarchical cognitive structures of intro physics problems

–References–

(1) M. T. H. Chi, P. J. Feltovich, and R. Glaser, *Cognitive Science* **5**, 121 – 152 (1981).

(2) T. de Jong, and M. G. Ferguson-Hessler, *Journal of Educational Psychology* **78**, 249 – 288 (1986).

(3) A. Mason, Master's thesis, University of Pittsburgh, Pittsburgh, PA, USA (2009).

(4) G. H. Veldhuis, *Science Education* **74**, 105 – 118 (1990).

(5) C. Singh, *American Journal of Physics* **77**, 73–80 (2009).

(6) S.-Y. Lin, and C. Singh, *European Journal of Physics* **31**, 57 (2010).