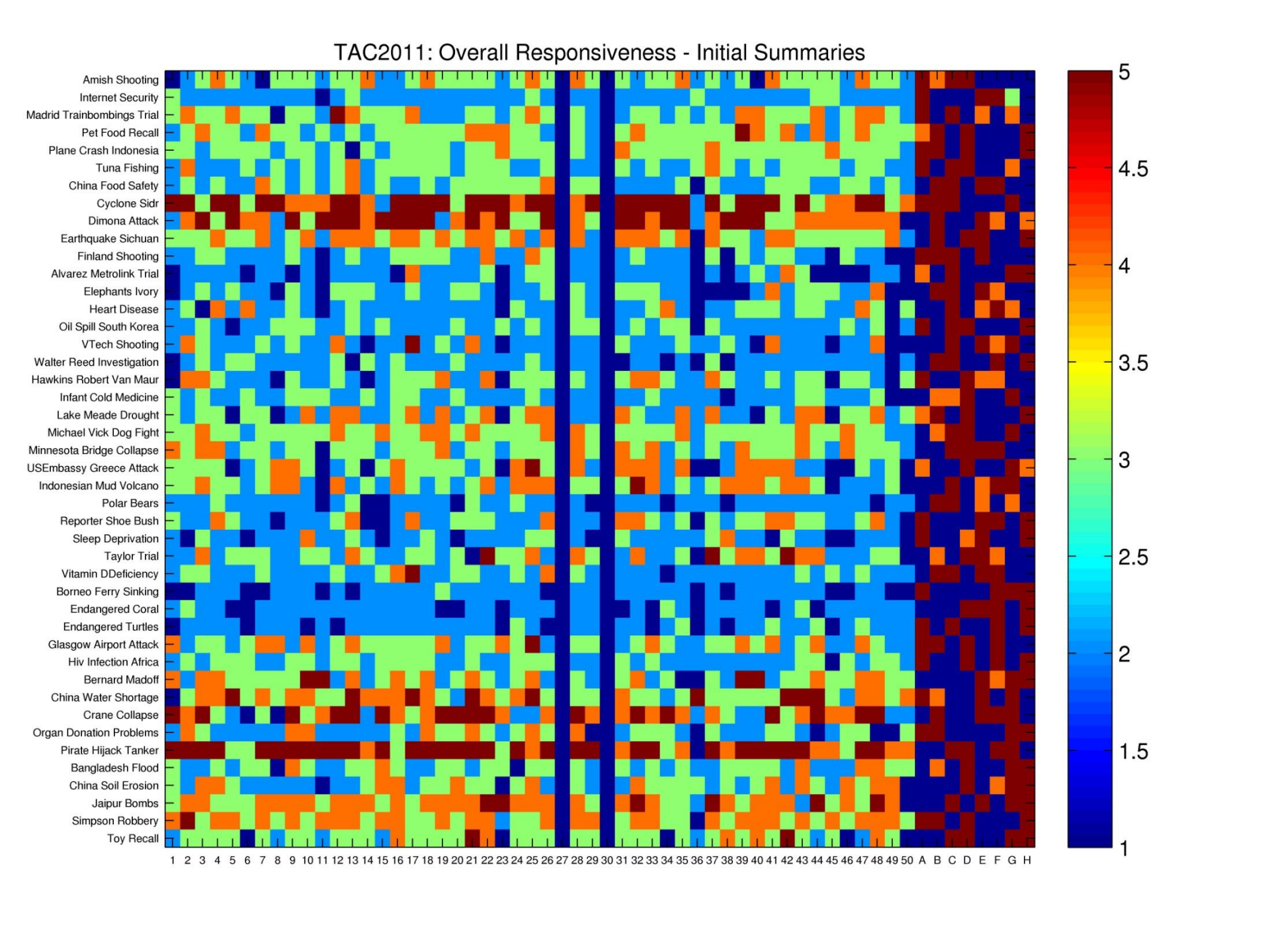
# Are these systems Significantly Different?

Peter A. Rankel, University of Maryland
John M. Conroy, IDA Center for Computing Sciences

November 14, 2011

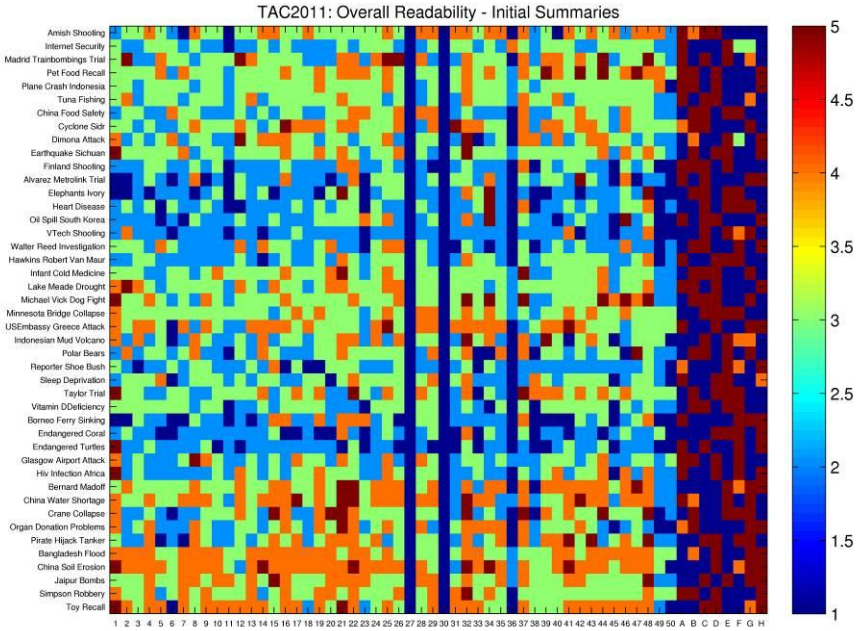
# Motivation

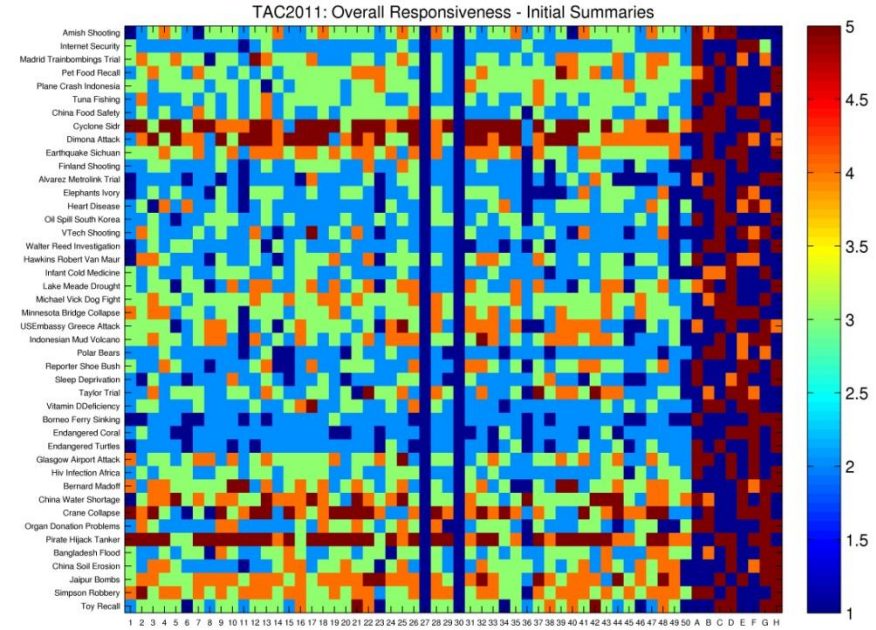
- Why do we propose using paired testing?
  - Paired testing can be more powerful than unpaired testing
  - Evidence suggests that document difficulty varies greatly, and paired testing accounts for this
  - Paired testing helps automatic metrics distinguish between humans and machines

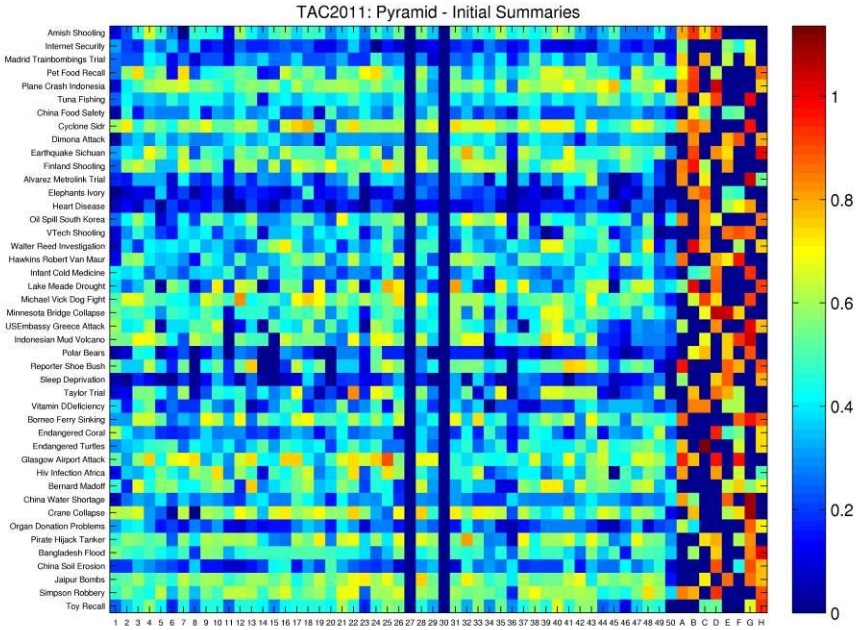

TAC2011: Overall Responsiveness - Initial Summaries
Amish Shooting
Internet Security
Madrid Trainbombings Trial
Pet Food Recall
Plane Crash Indonesia
Tuna Fishing
China Food Safety
Cyclone Sidr
Dimona Attack
Earthquake Sichuan
Finland Shooting
Alvarez Metrolink Trial
Elephants Ivory
Heart Disease
Oil Spill South Korea
VTech Shooting
Walter Reed Investigation
Hawkins Robert Van Maur
Infant Cold Medicine
Lake Meade Drought
Michael Vick Dog Fight
Minnesota Bridge Collapse
USEmbassy Greece Attack
Indonesian Mud Volcano
Polar Bears
Reporter Shoe Bush
Sleep Deprivation
Taylor Trial
Vitamin DDeficiency
Borneo Ferry Sinking
Endangered Coral
Endangered Turtles
Glasgow Airport Attack
Hiv Infection Africa
Bernard Madoff
China Water Shortage
Crane Collapse
Organ Donation Problems
Pirate Hijack Tanker
Bangladesh Flood
China Soil Erosion
Jaipur Bombs
Simpson Robbery
Toy Recall
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 A B C D E F G H
5
4.5
4
3.5
3
2.5
2
1.5
1

Initial Summaries:

The A Set

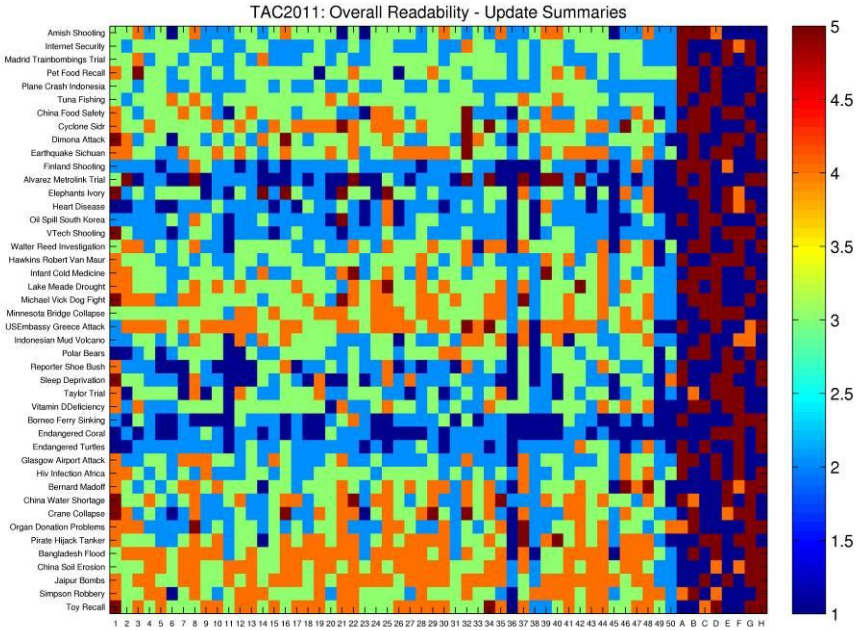


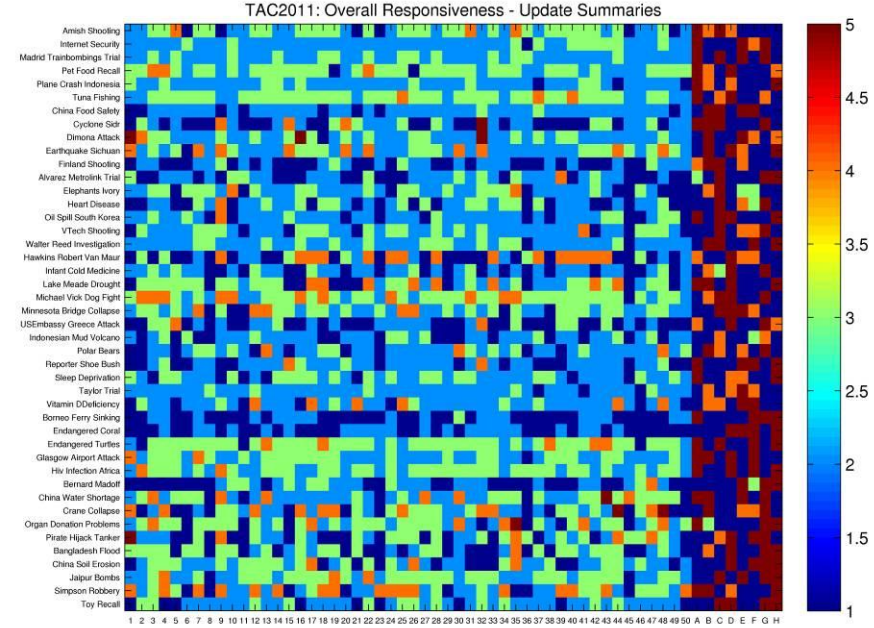



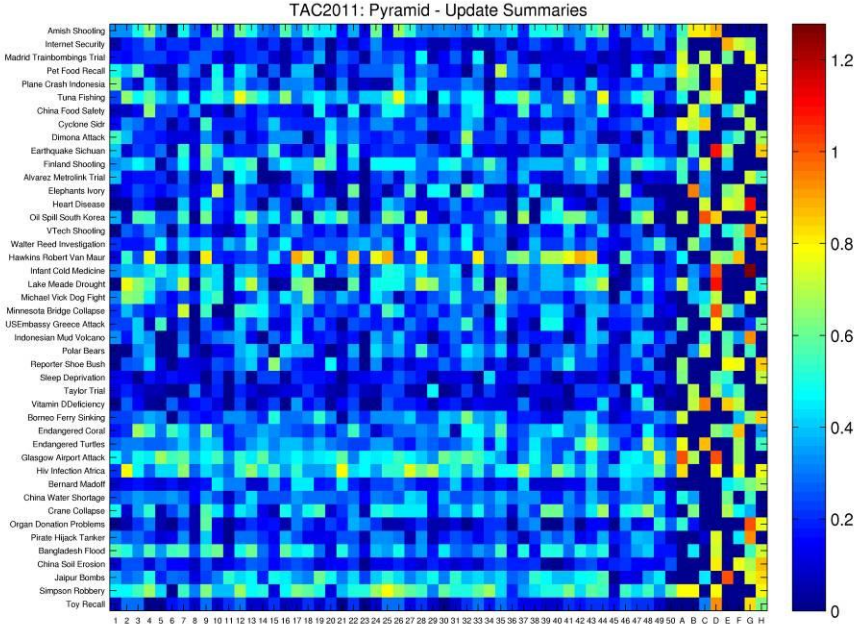



Update Summaries:

The B Set

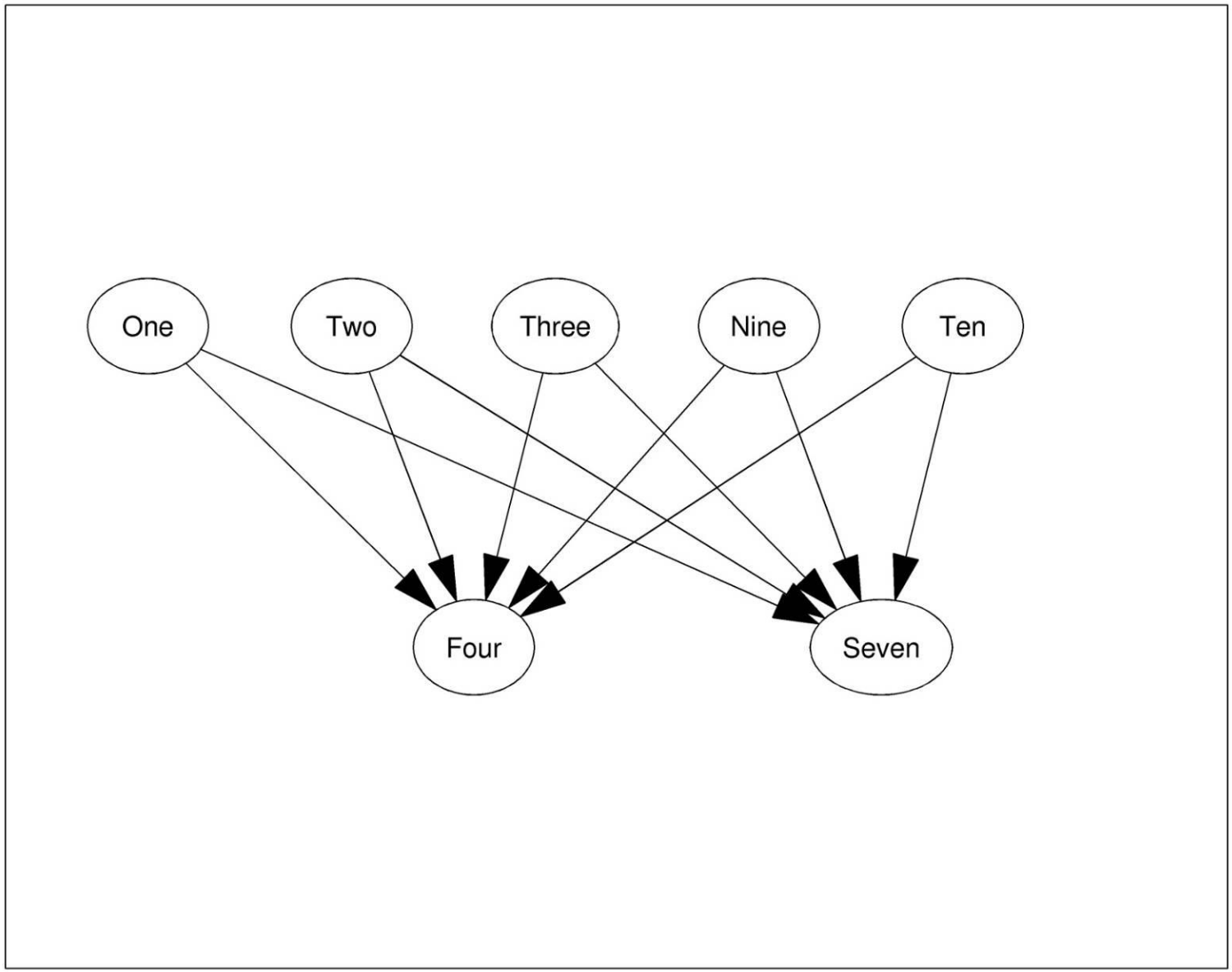


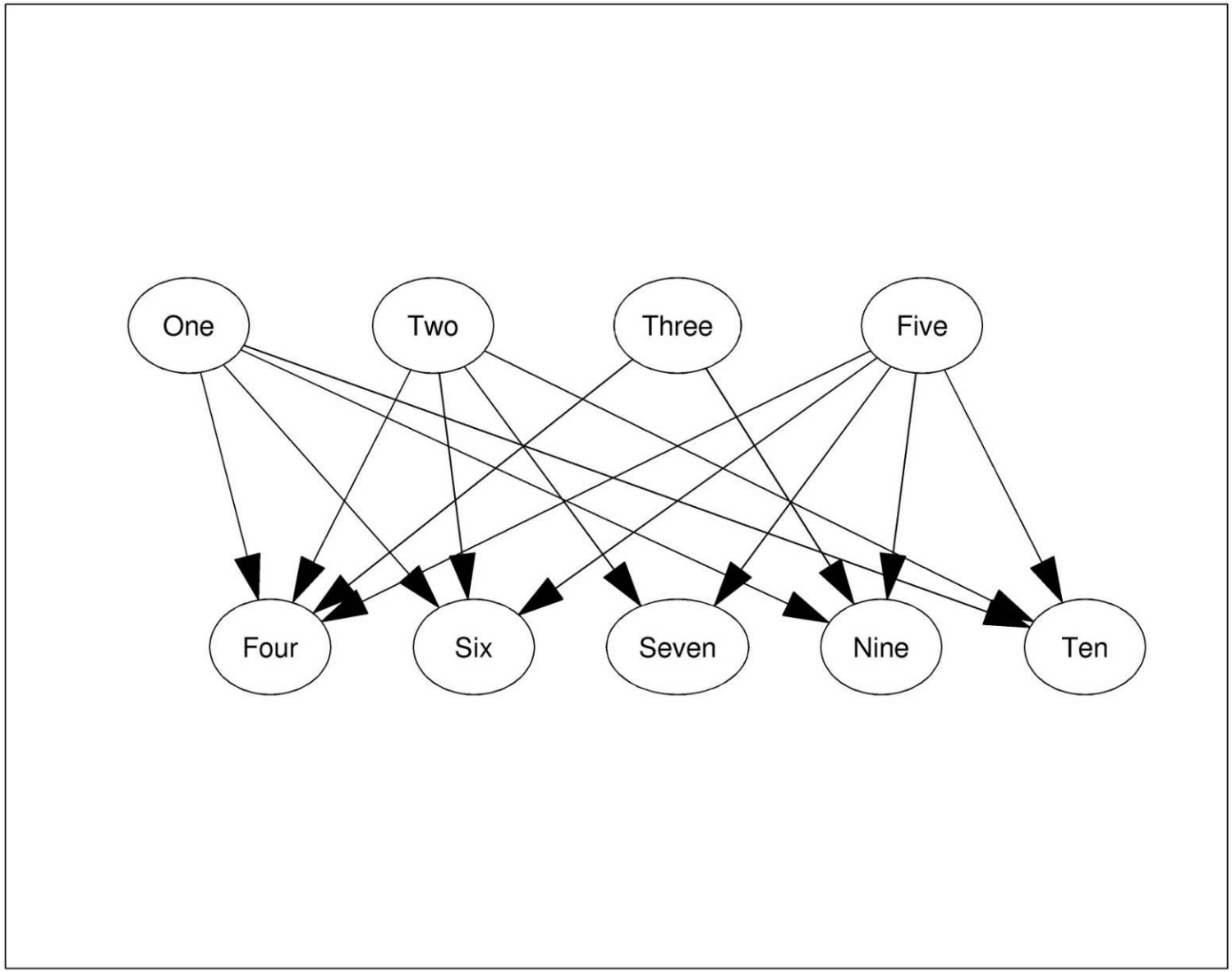



# MultiLingual Task

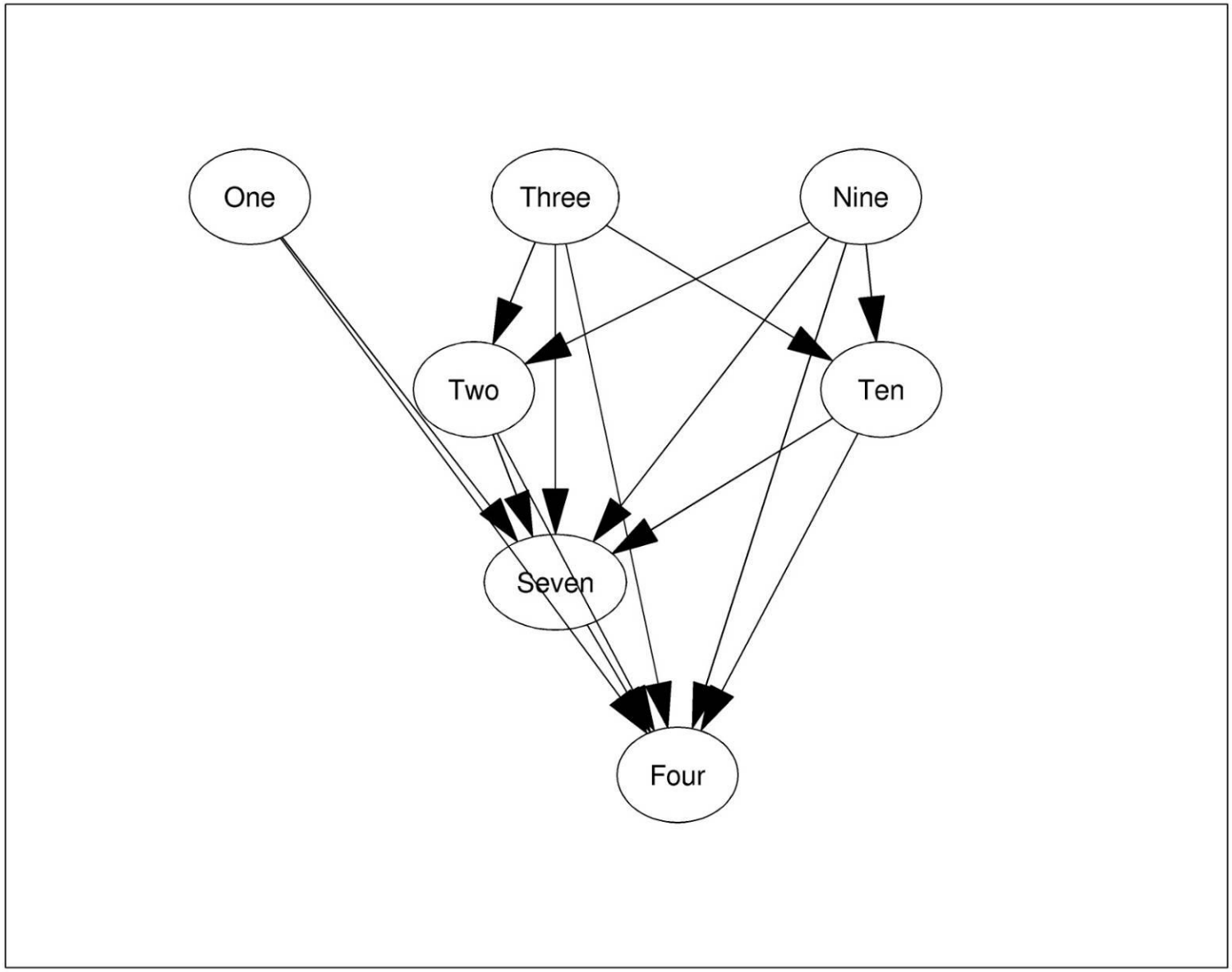
- Use paired testing to evaluate system performance on individual languages
- Draw a directed graph displaying an edge (i,j) only when system i significantly outperformed system j
- Paired test was the non-parametric Wilcoxon signed-rank test
- Used Bonferroni adjustment to account for multiple comparisons

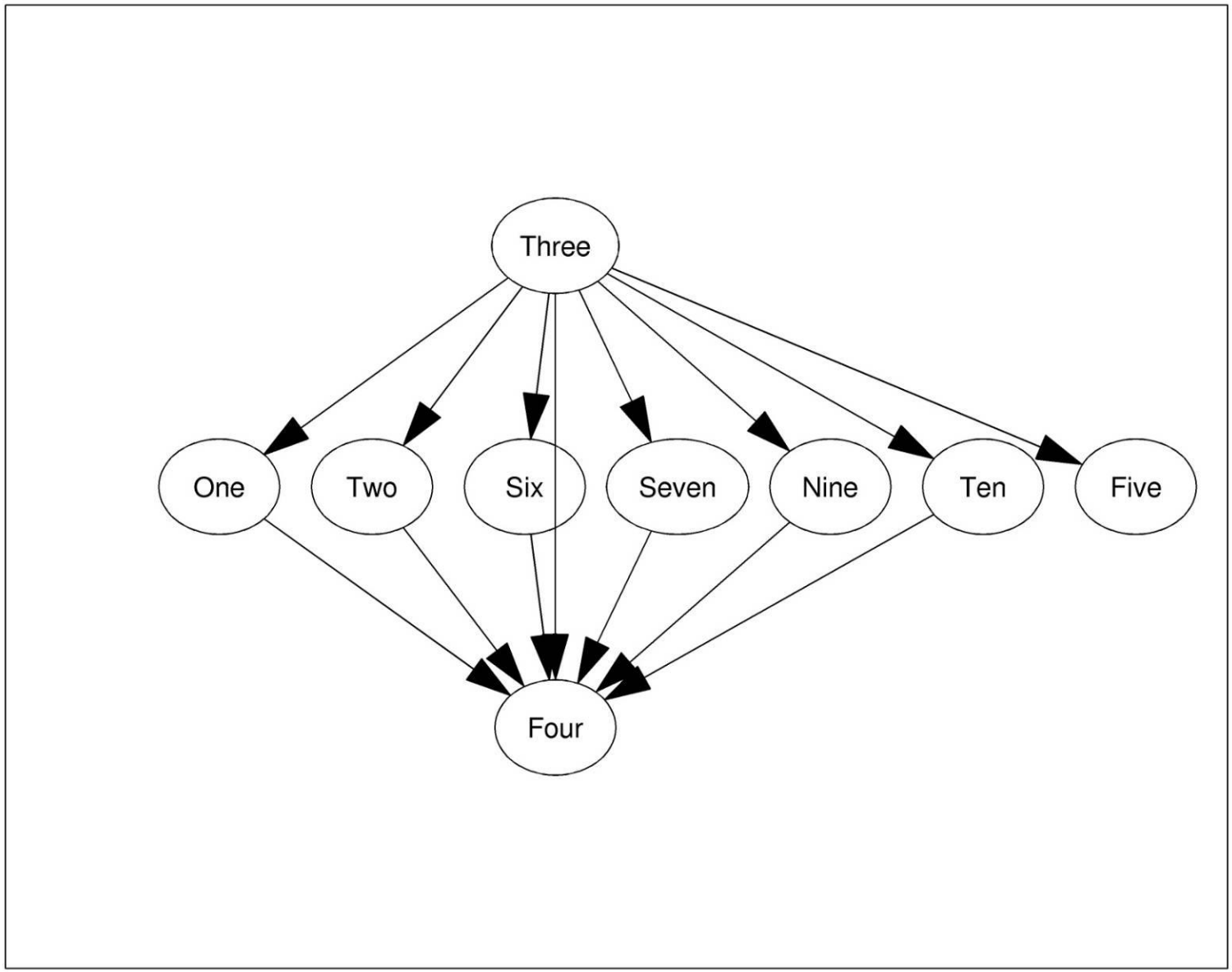

greek
One
Two
Three
Nine
Ten
Four
Seven

hindi

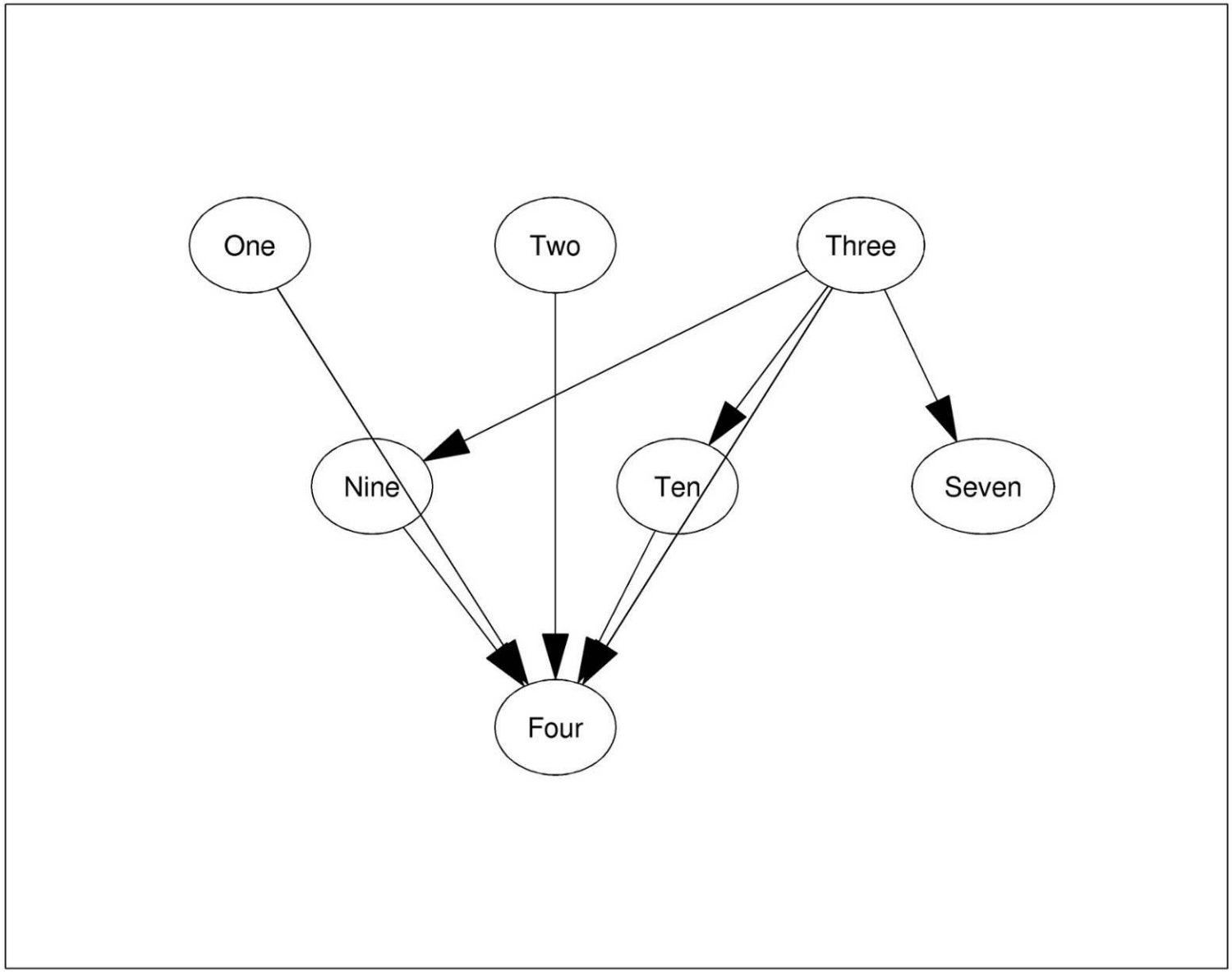


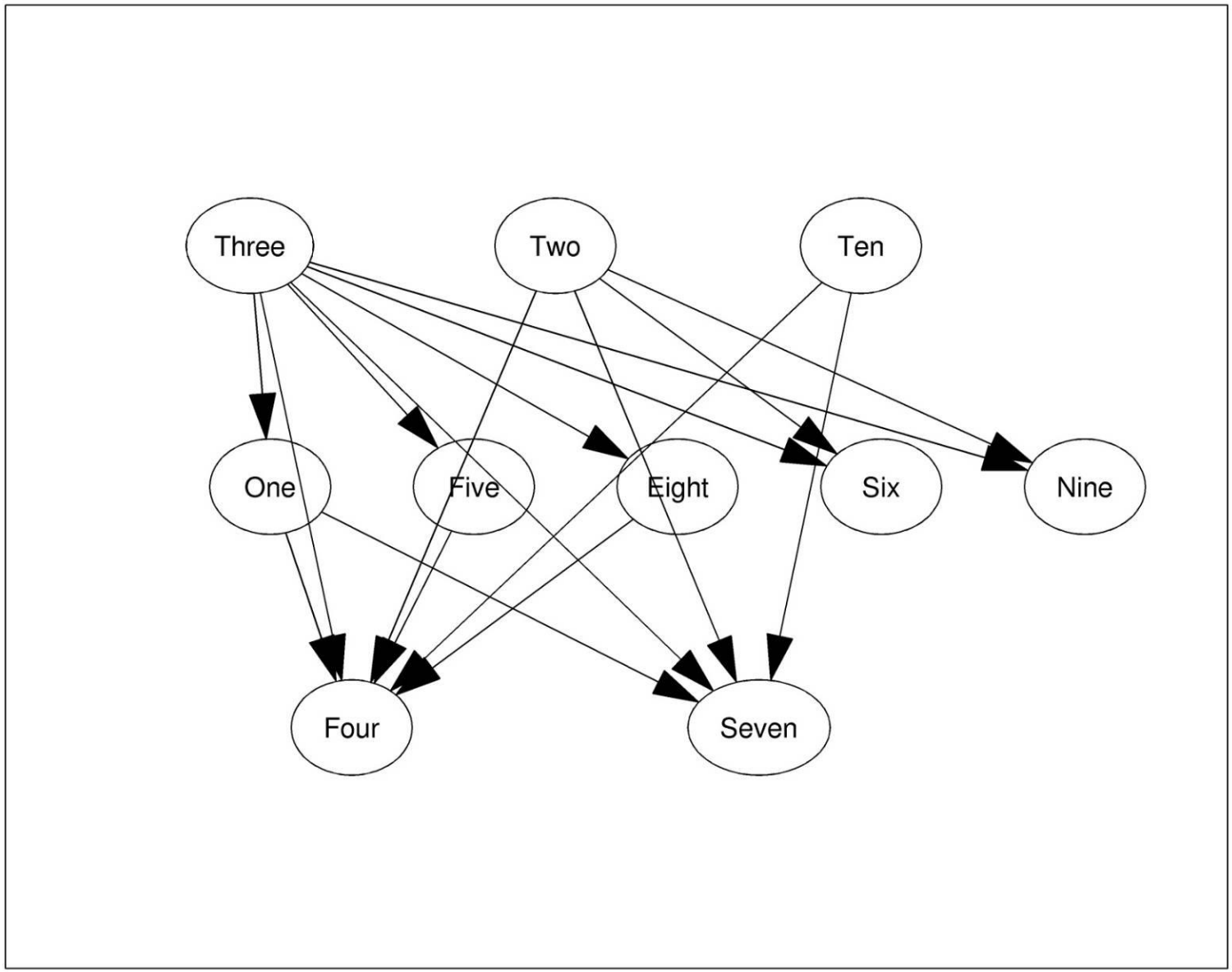
czech

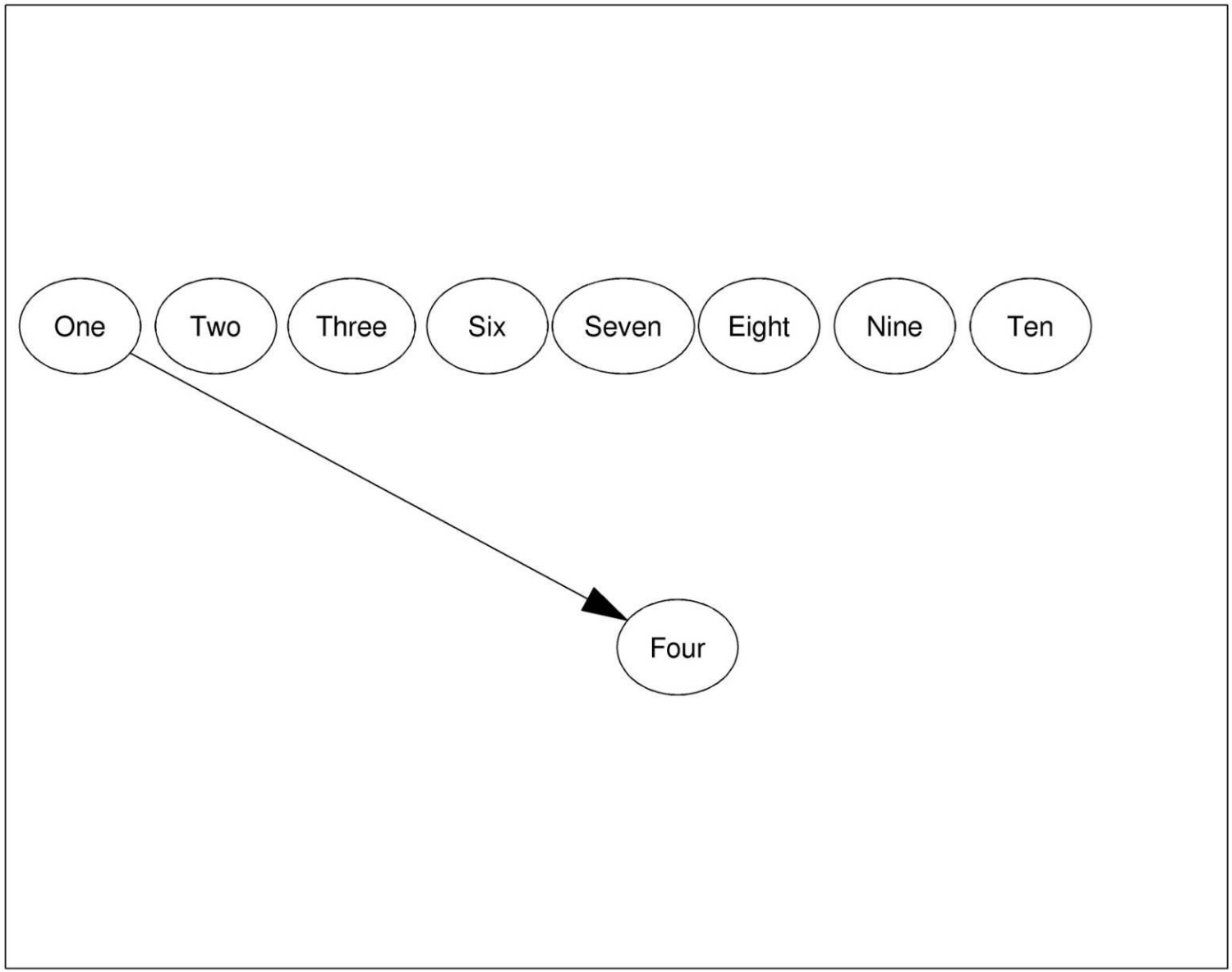



french
Three
One
Two
Six
Seven
Nine
Ten
Five
Four

hebrew

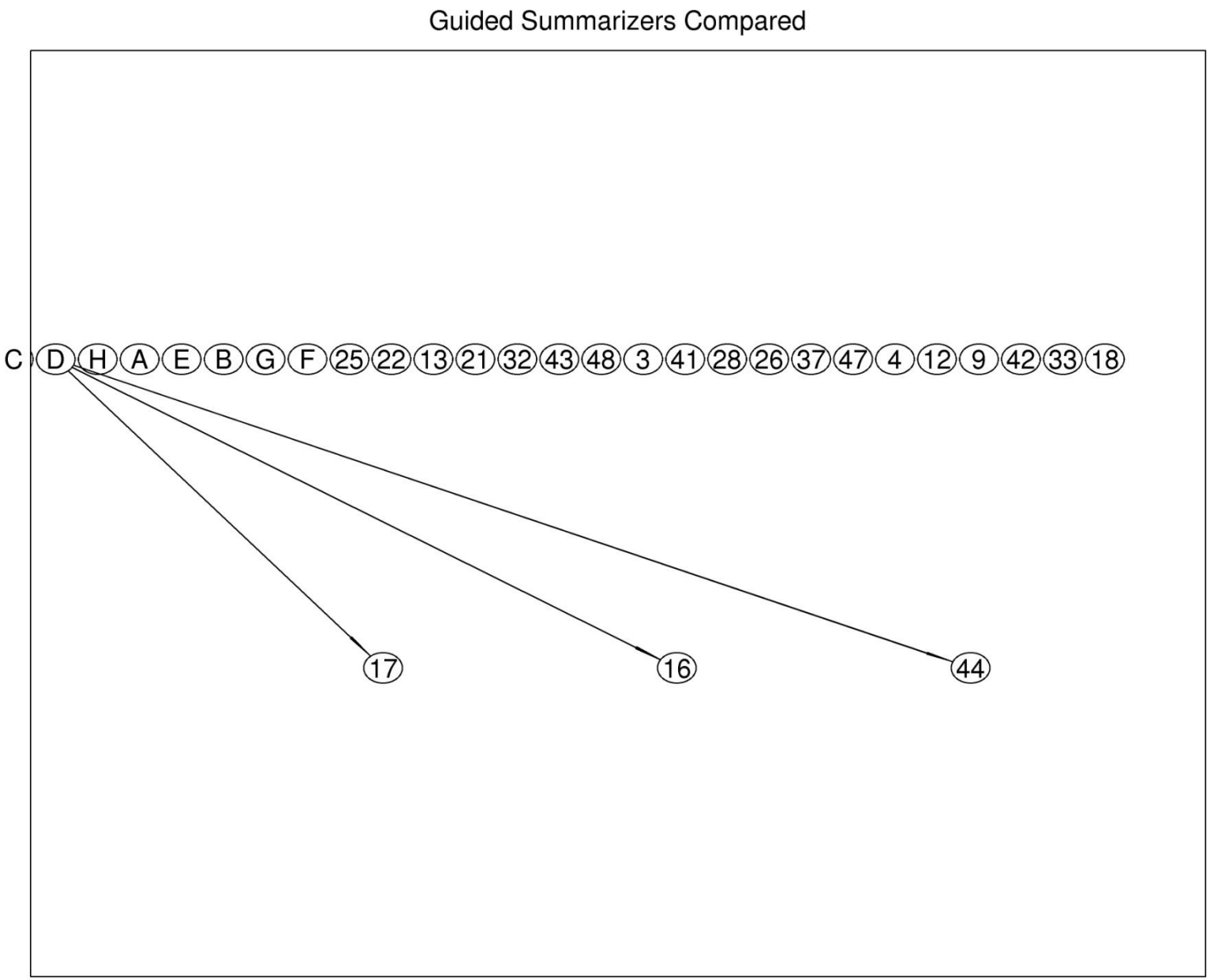



english
Three
Two
Ten
One
Five
Eight
Six
Nine
Four
Seven


arabic
One
Two
Three
Six
Seven
Eight
Nine
Ten
Four

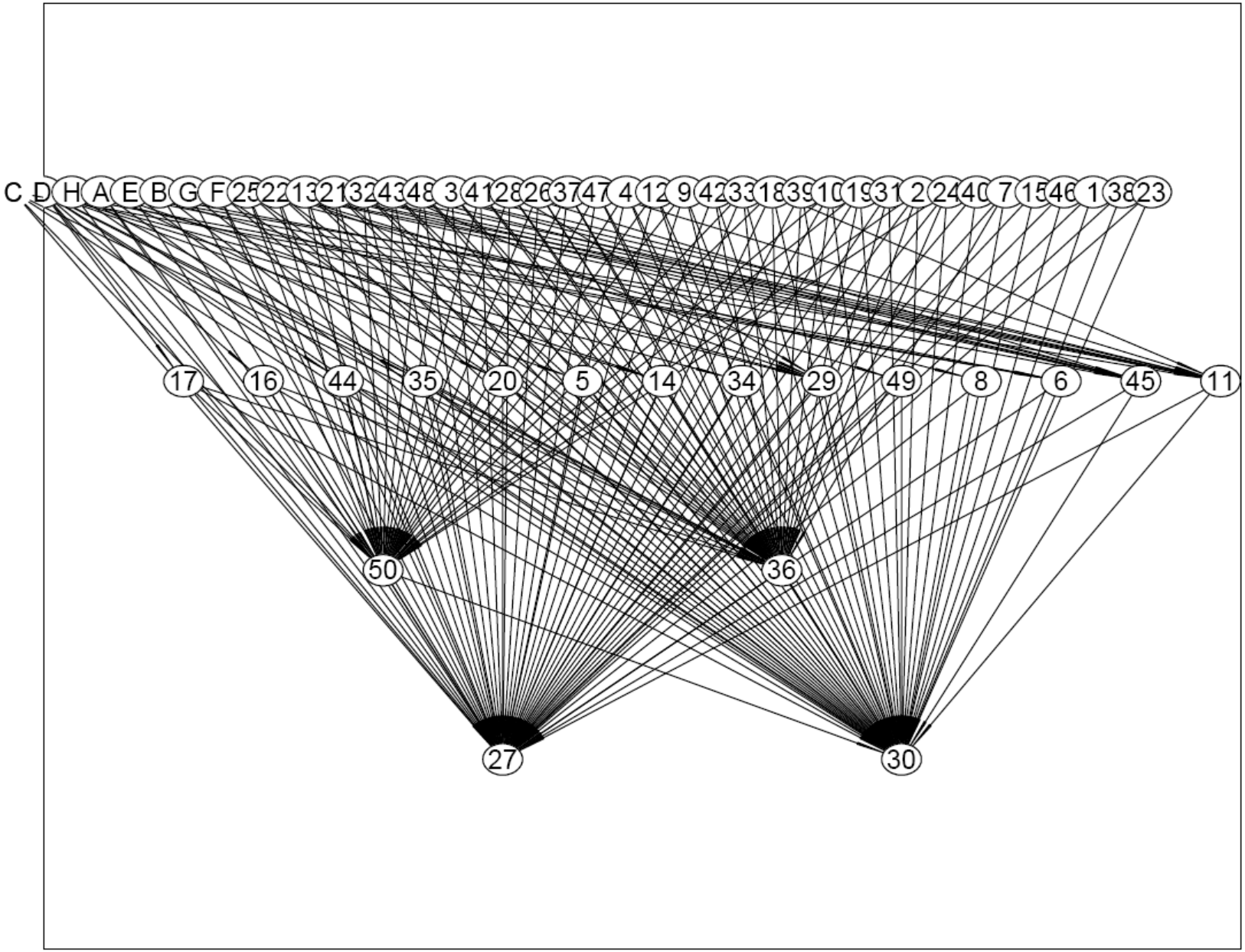
# Guided Summarization

- Apply the same idea to systems in the Guided Summarization task
- Systems 1-50
- Humans A-H
- Wilcoxon signed-rank test
- Bonferroni Correction




The Top 30 Summarization Systems

Guided Summarizers Compared



# Future Work

- More closely examine the AESOP results
  - Re-calculate the tables of discriminative significant differences using paired testing
- Look at TAC and DUC data from past years to see if there are trends.
  - E.g. Does it get more crowded at the top for a new task or a repeated task?

# References


- Peter Rankel, John M. Conroy, Eric V. Slud, and Dianne P. O'Leary, “Ranking Human and Machine Summarization Systems,” Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP 2011), Edinburgh, UK, July 27-31, 2011. Association for Computational Linguistics (ACL) (link to pdf)