Supplementary Material: Bayesian Checking for Topic Models

David Mimno
Department of Computer Science
Princeton University Princeton, NJ 08540
`mimno@cs.princeton.edu`

David Blei
Department of Computer Science
Princeton University Princeton, NJ 08540
`blei@cs.princeton.edu`

1 Topic Visualization

Nine topics (three low variability, three medium, and three high) are shown for the CMU 2008 Political Blogs corpus (Figure 1) and British Parliament corpus (Figure 2).

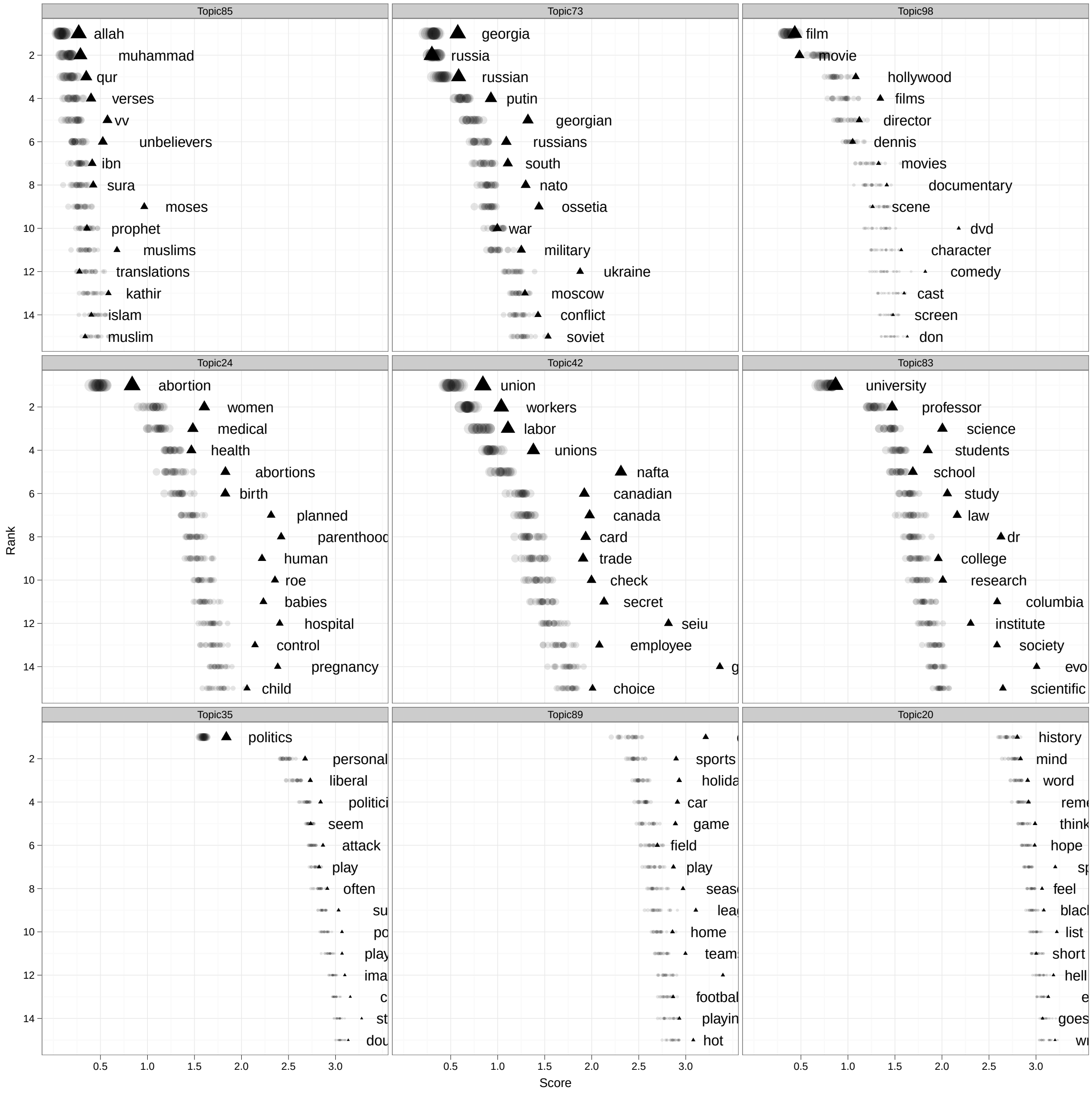

Figure 1: **Visualization of variability within topics.** Nine randomly selected topics from the CMU Political Blog corpus with low (top row), medium (middle row) and high (bottom row) mutual information between words and documents. The y-axis shows term rank within the topic, with size proportional to log probability. The x-axis represents divergence from the multinomial assumption for each word: terms that are uniformly distributed across documents are towards the left, while more specialized terms are to the right. Triangles represent real values, circles represent 20 replications of this same plot from the posterior of the model. Size is proportional to log probability.

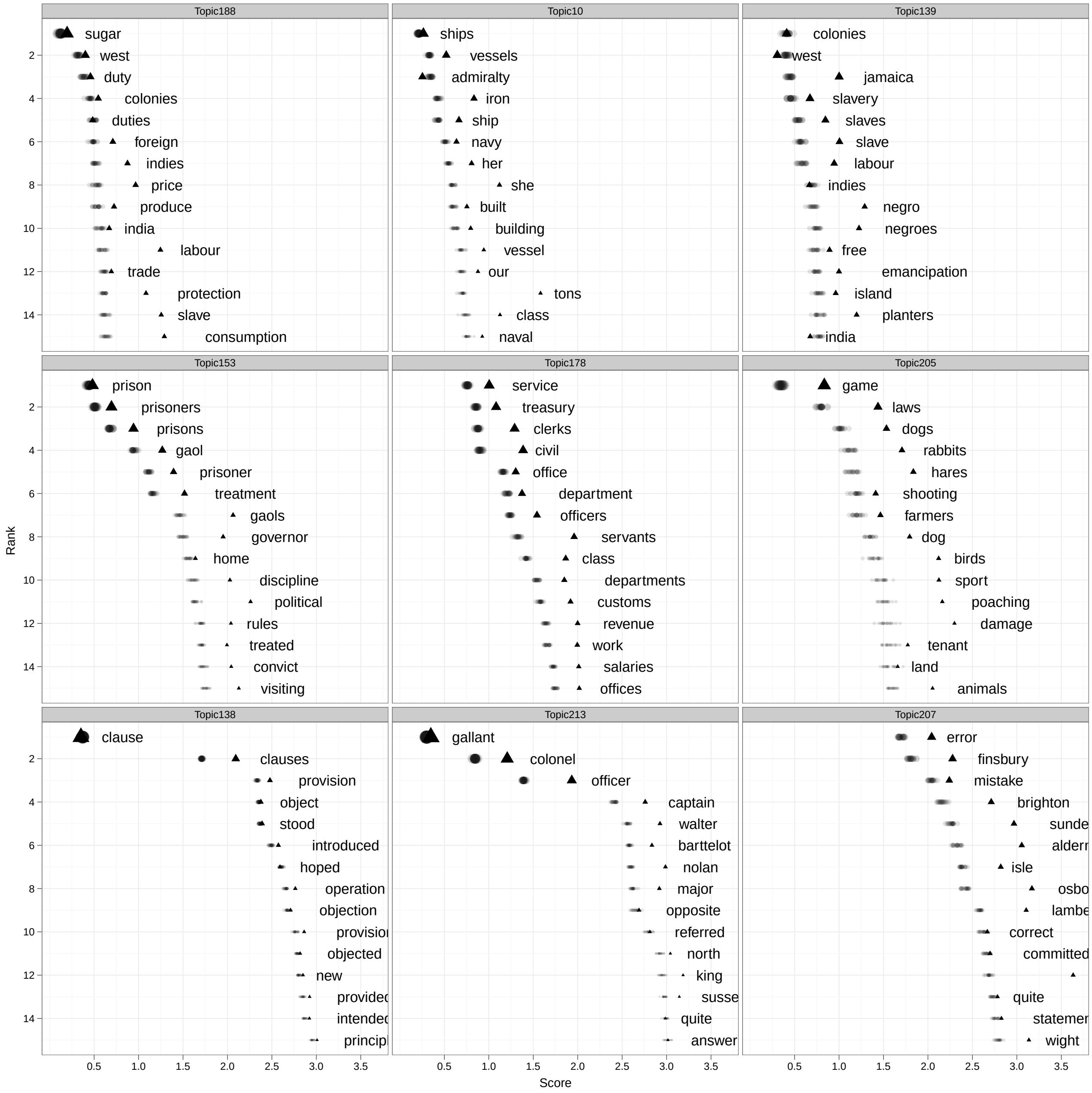

Figure 2: **Visualization of variability within topics.** Nine randomly selected topics from the Parliament corpus with low (top row), medium (middle row) and high (bottom row) mutual information between words and documents. The y-axis shows term rank within the topic, with size proportional to log probability. The x-axis represents divergence from the multinomial assumption for each word: terms that are uniformly distributed across documents are towards the left, while more specialized terms are to the right. Triangles represent real values, circles represent 20 replications of this same plot from the posterior of the model. Size is proportional to log probability.