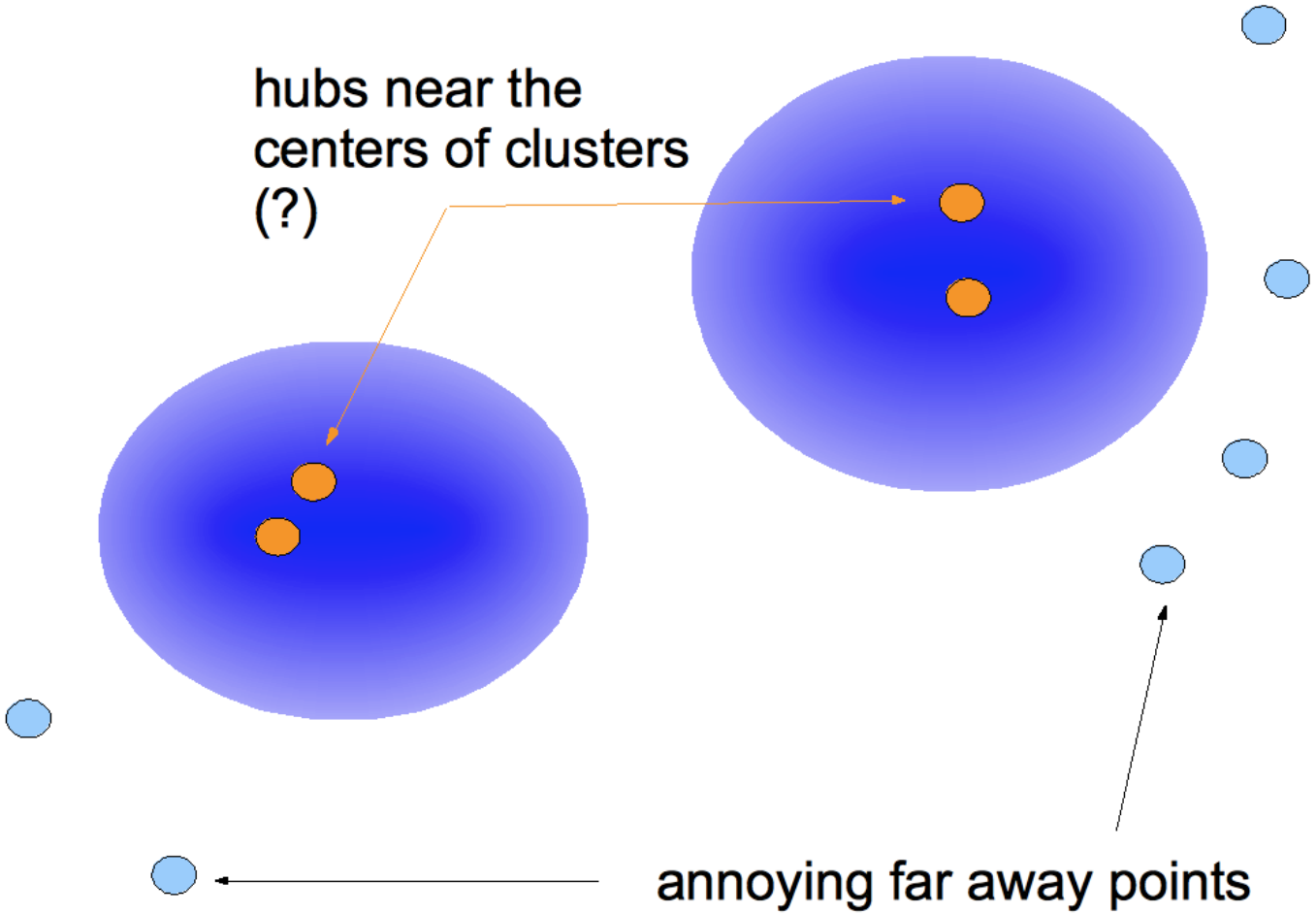
Density and Hubness

Libby Beer (libby.beer@gmail.com)
Jesse Metcalf-Burton (datasciencejess@gmail.com)
Marilyn Vazquez (mvazque3@masonlive.gmu.edu)

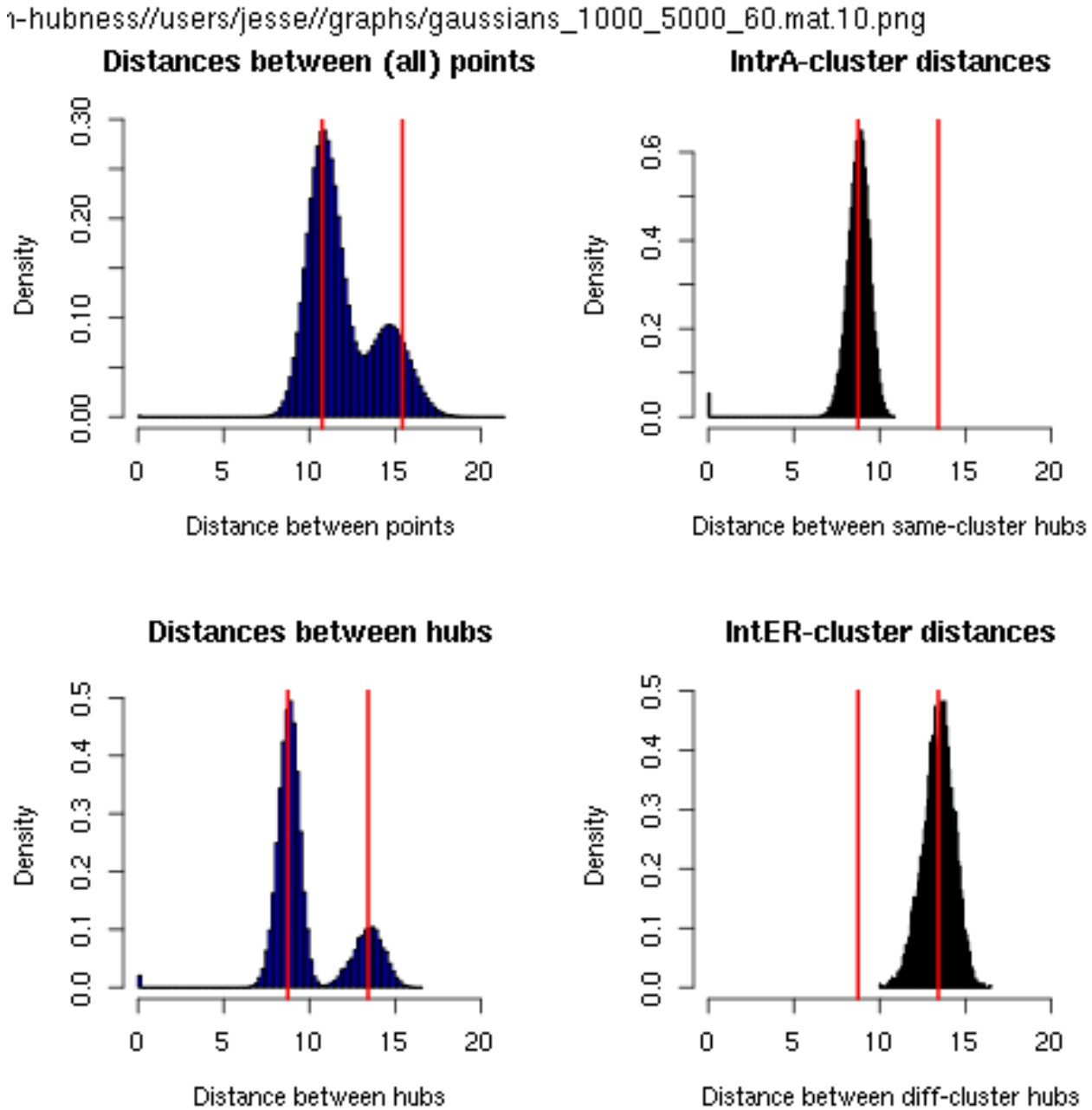
July 21, 2017

Distances between points? Hubs?

Can hubs help us cluster?

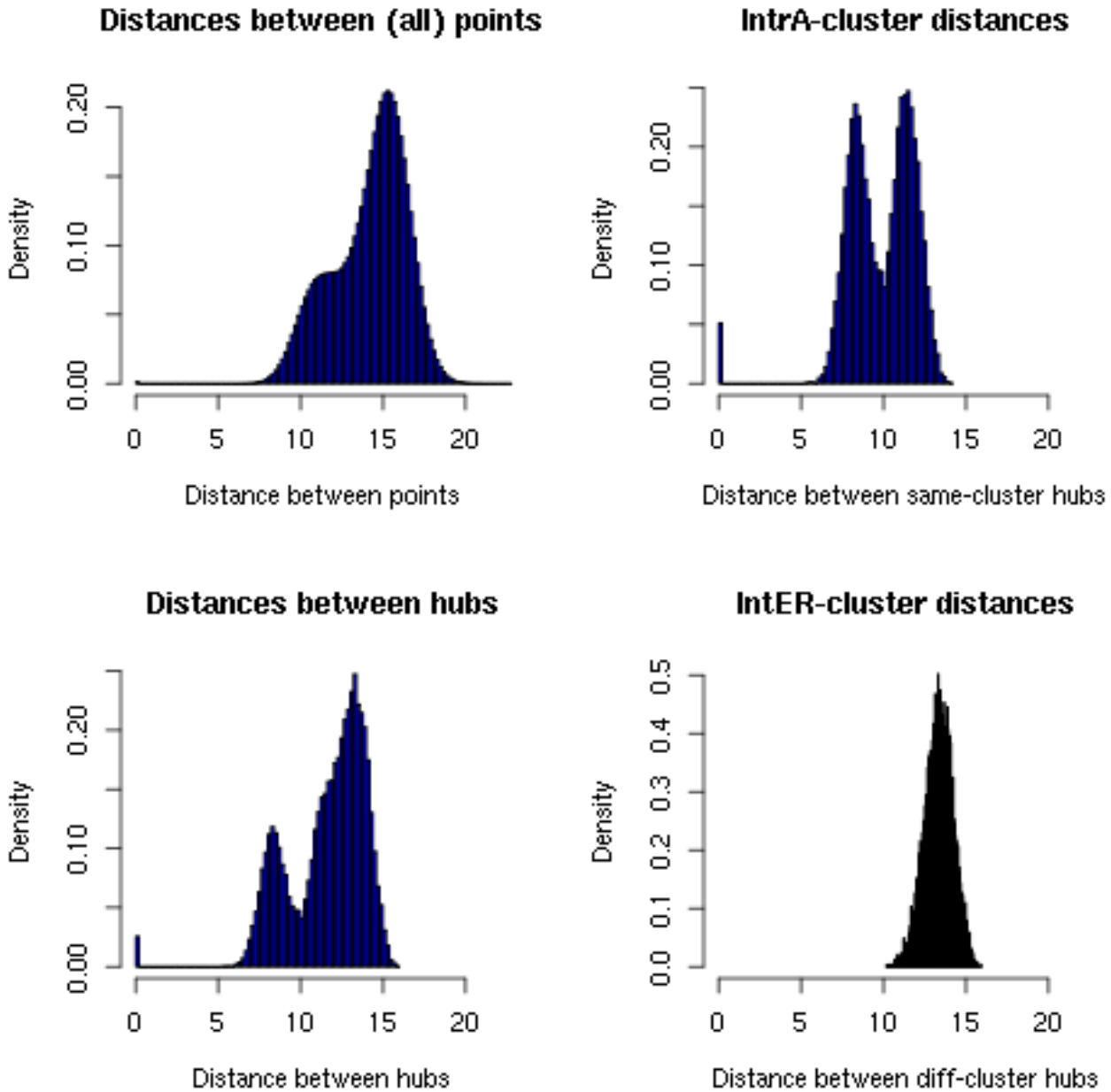


Distances with synthetic data



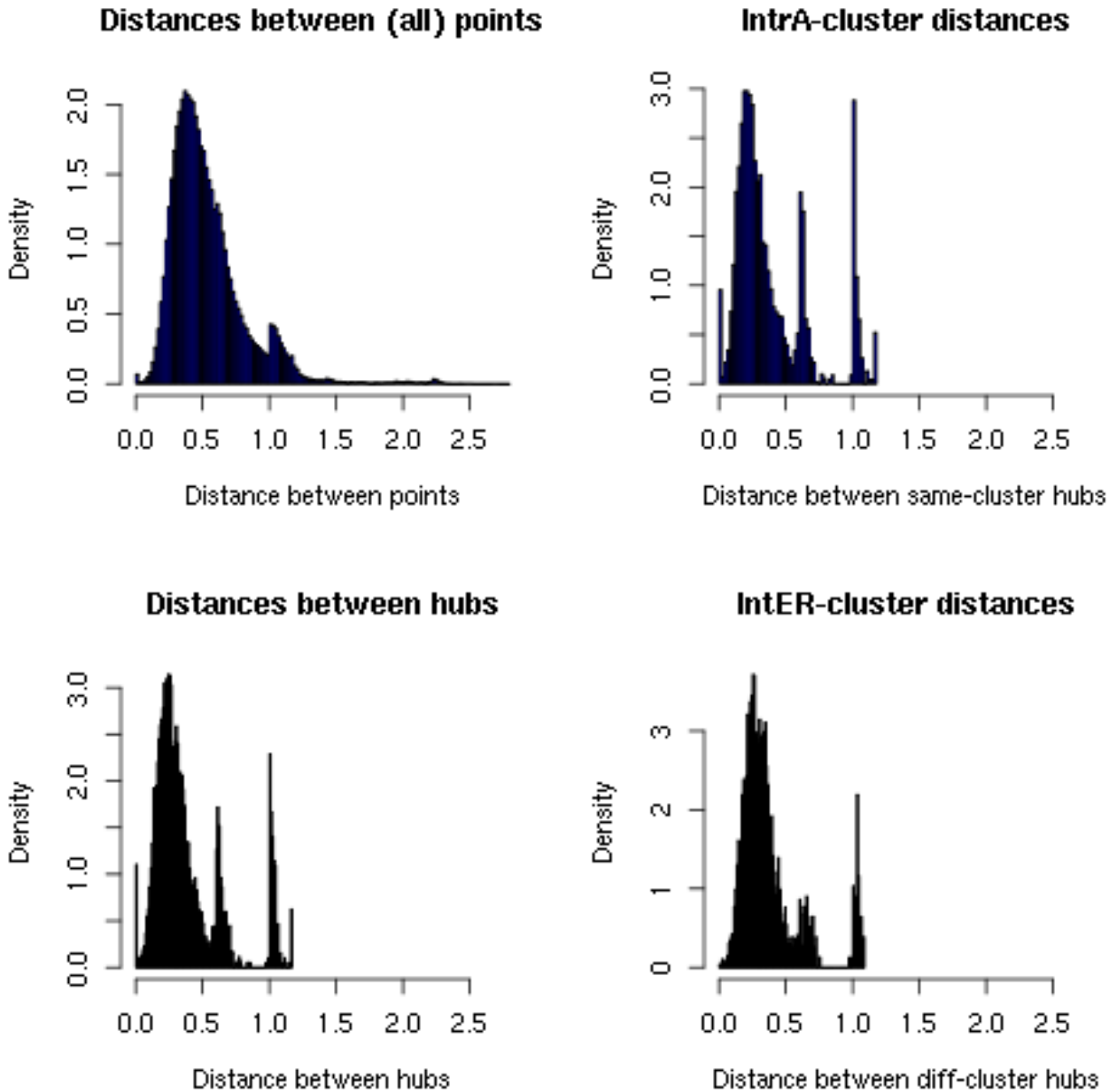
Distances with (other) synthetic data

ıbness//users/jesse//graphs/synthetic_data_30_50_overlap30.mat.10.png



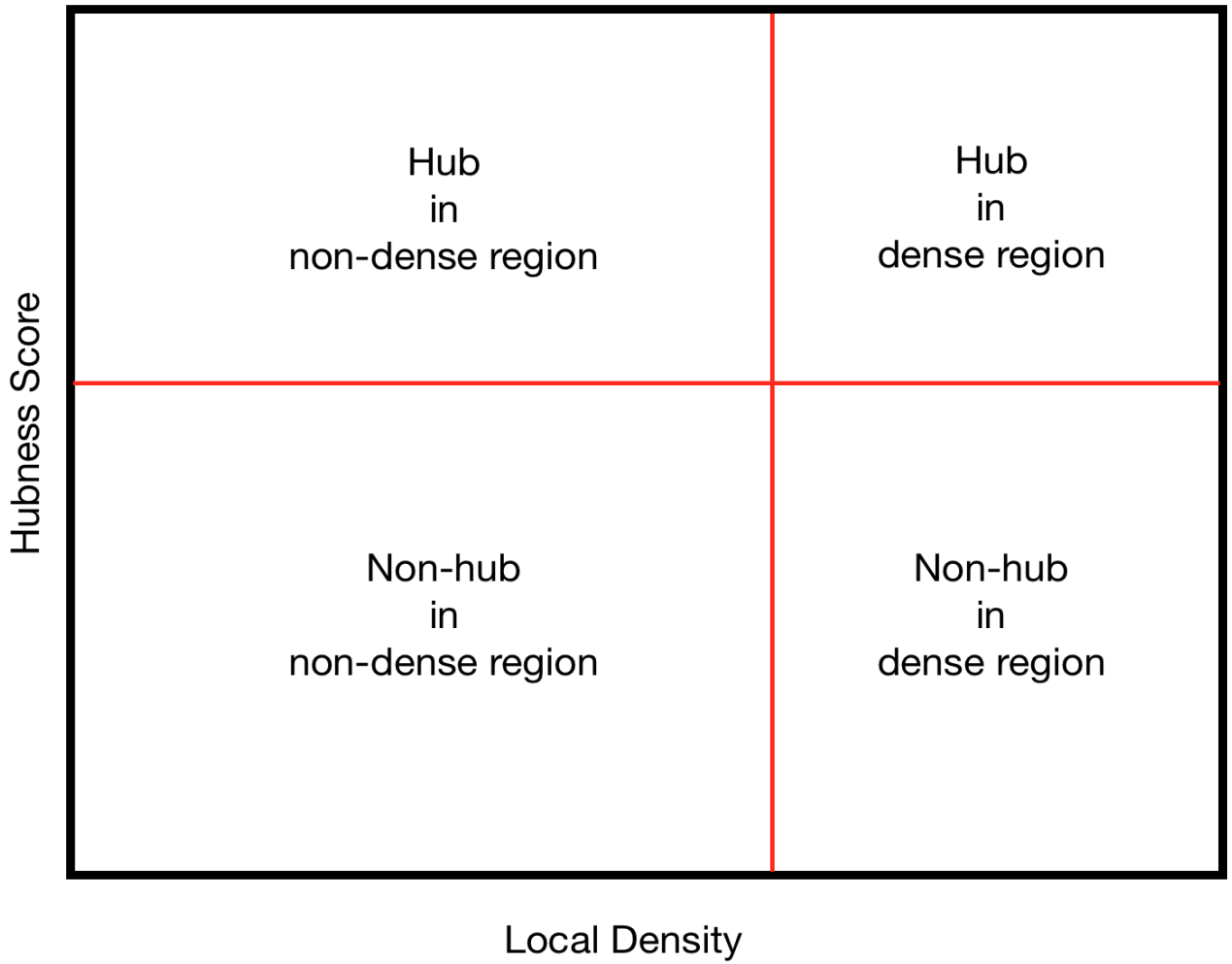
Distances with real data

:7/wisdm-hubness//users/jesse//graphs/data_spam.mat.10.png



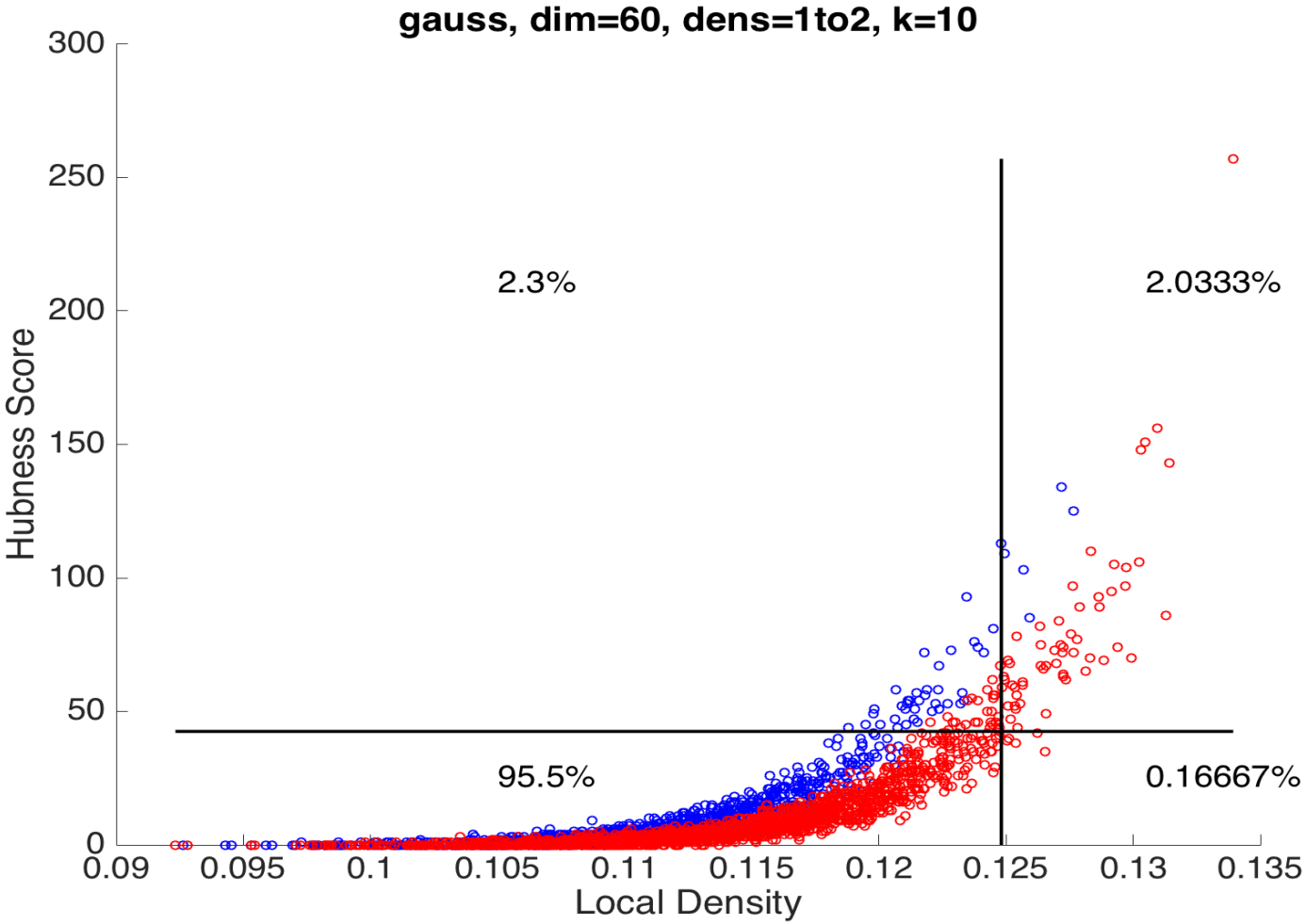
Do hubs appear in high density regions?

Intuition



Do hubs appear in high density regions?

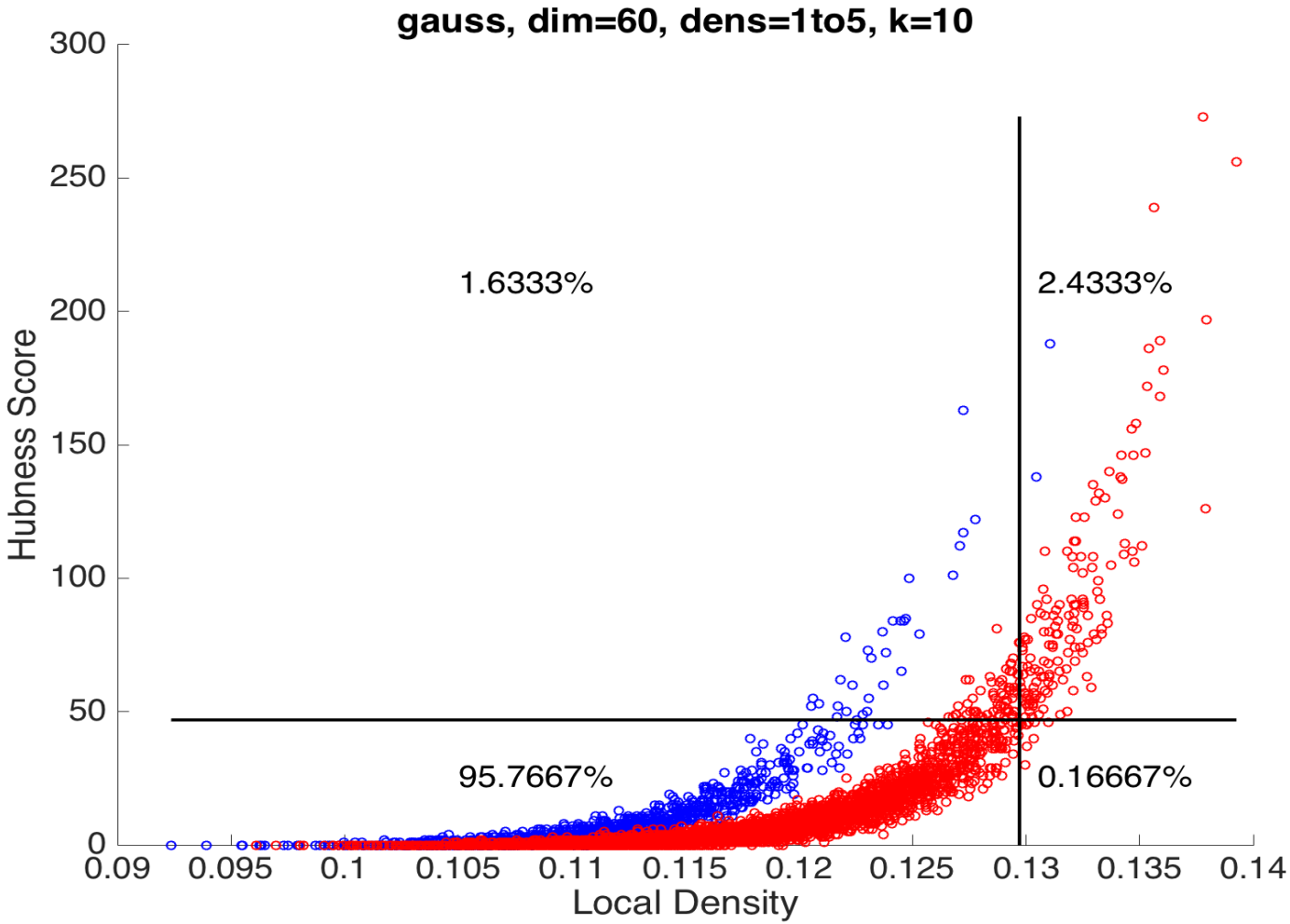
Answer: Sometimes!



Notice: (1) Proportion of hubs in the two density regions and (2) proportion of high density points that are hubs.

Do hubs appear in high density regions?

Answer: Sometimes!



Notice: (1) A LOT more hubs in the denser cluster and (2) the rise of two "wings" that correspond to the two different clusters.

Observations

From the experiments on the Gaussian data, we observe that:

- About 13 out of 36 experiments had a larger proportion of hubs in the low density region: hubness scores and density are not exactly the same thing.
- About 23 out of 36 experiments had a larger proportion of hubs in the high density region: hubs are more likely to be in a high density region.
- All experiments show that high density regions are mostly composed of hubs: if a point is in a high density region, it is most likely a hub.

Future Work

- More experiments of local density versus hubness scores for other density types e.g. uniform density
- Verify if patterns also show up in real data
- Data with more clusters to see if "wings" arise for each cluster
- Test model selection via clustering only hubs
- Comparison of hubs to core points, outliers, etc.