Topic Modeling and Insights from AITA Subreddit

Henry Arze[†], and Paras Goda[†]

RESEARCH QUESTION

How can we:
- identify topics from the AITA subreddit using transformer-based topic modeling
- measure the influence of these topics on contextual meaning in moral language
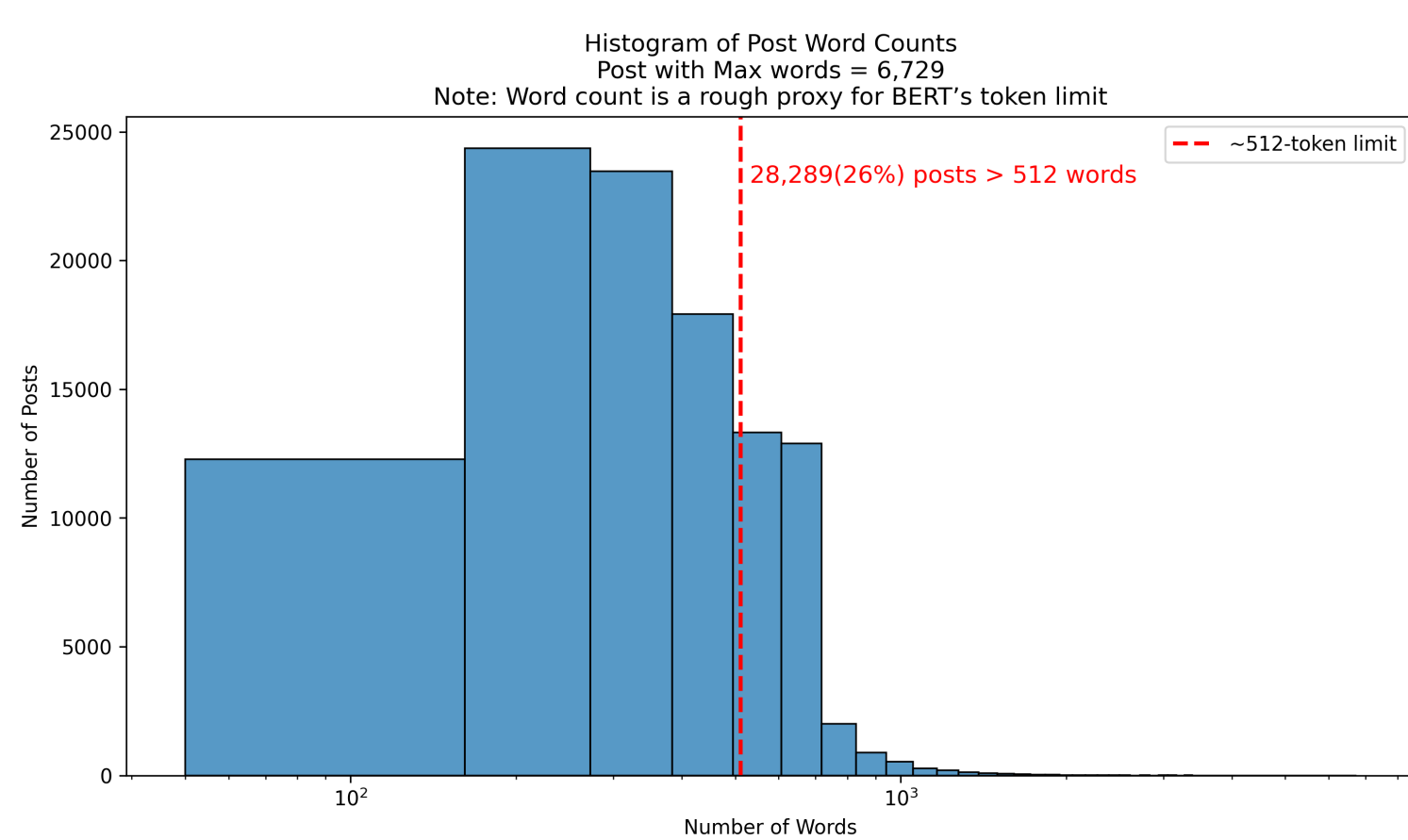
MOTIVATION

- AITA posts contain diverse social and moral conflict
- Transformer embeddings can produce more coherent, contextually rich topics
- Understanding topics and topic pairs helps show how moral language shifts in different scenarios

PREVIOUS STUDY

Nguyen et al. [1] performed a topic modeling analysis on AITA by:
- Carefully computing optimal k for k-means
- Emphasizing the advantages for topic pairs by showing the shifts in moral language

PRELIMINARY EDA

Posts with 50+ word and 10 comments

TOPIC MODELING STEPS

MPNet Clustering

- Generate doc embeddings (MPNet)
- Perform dimension reduction (UMAP)
- Cluster semantically similar posts (HDBSCAN)
- Output most representative words (c-TF-IDF)

Model	Coherence (c_v)	# Topics
MPNet-base	0.6023	48
RoBERTa-base	0.5466	62

Topic quality across transformer models.

AITA SUBREDDIT THREADS

Topic: formal relationship & depression

My SO and I have been dating for about 4 1/2 years now yet we barely have any physical intimacy. [...]
Aside from that detail she's quite possibly the closest thing to a superhuman I've ever seen. She runs her own business, helps me maintain the home, **destroys** me in smash, makes me laugh until my sides hurt, frequently buys me gifts and listens to me vent about work. [...]

So my girlfriend (whom I live with) has always wanted a dog. [...] Well fast forward a few months later and here we are with a 60lb 8 month old boxer... And I hate it. ... When left alone, It chews anything near by, **destroys** blinds, tears up carpet and generally trashes the place.
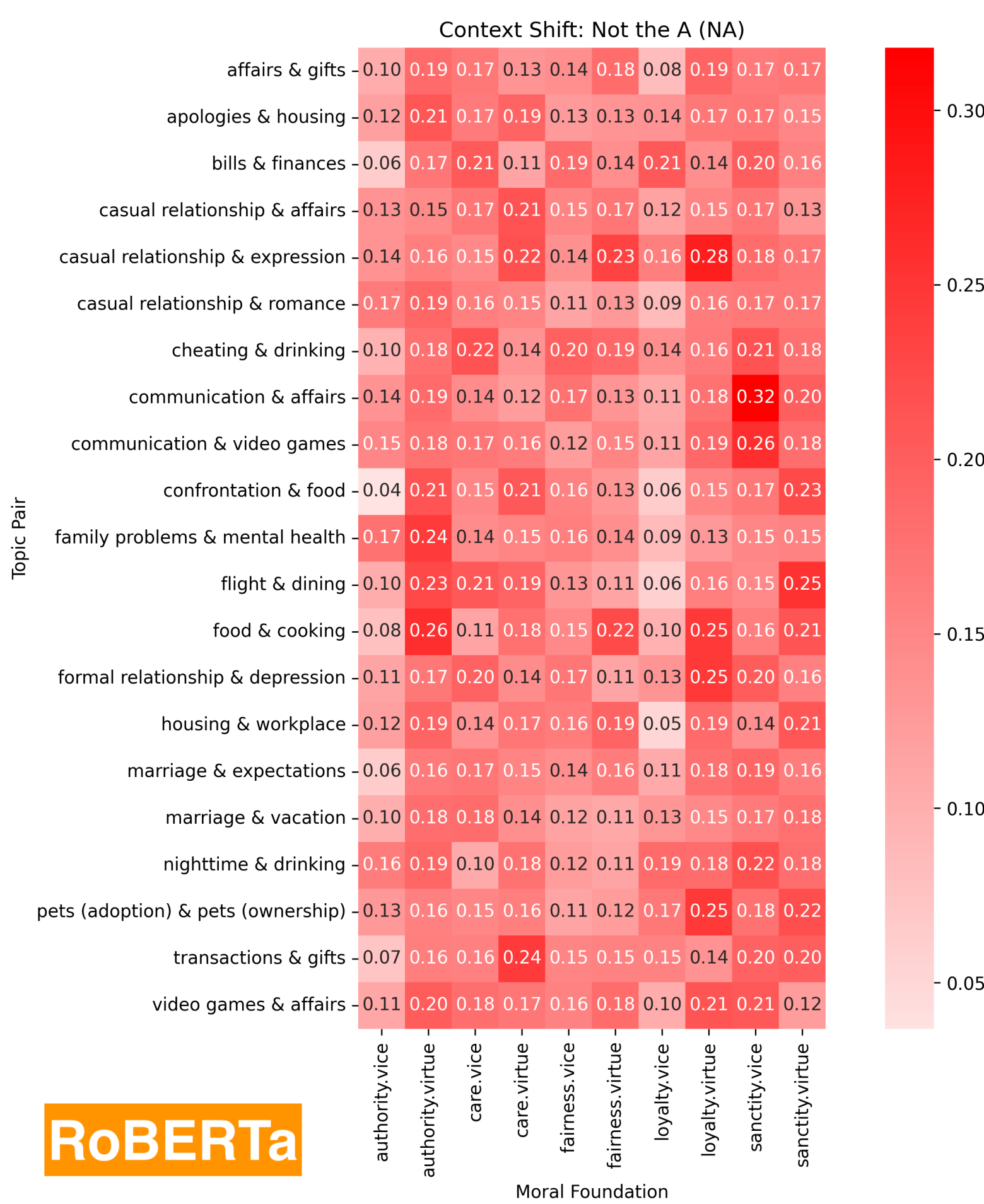
Topic: marriage & expectations

So basic background is I'm 42, two kids age 14 and 12 and signed divorce papers in April after almost 2 years of it being **drug** out trying to get my fair share from him. [...]
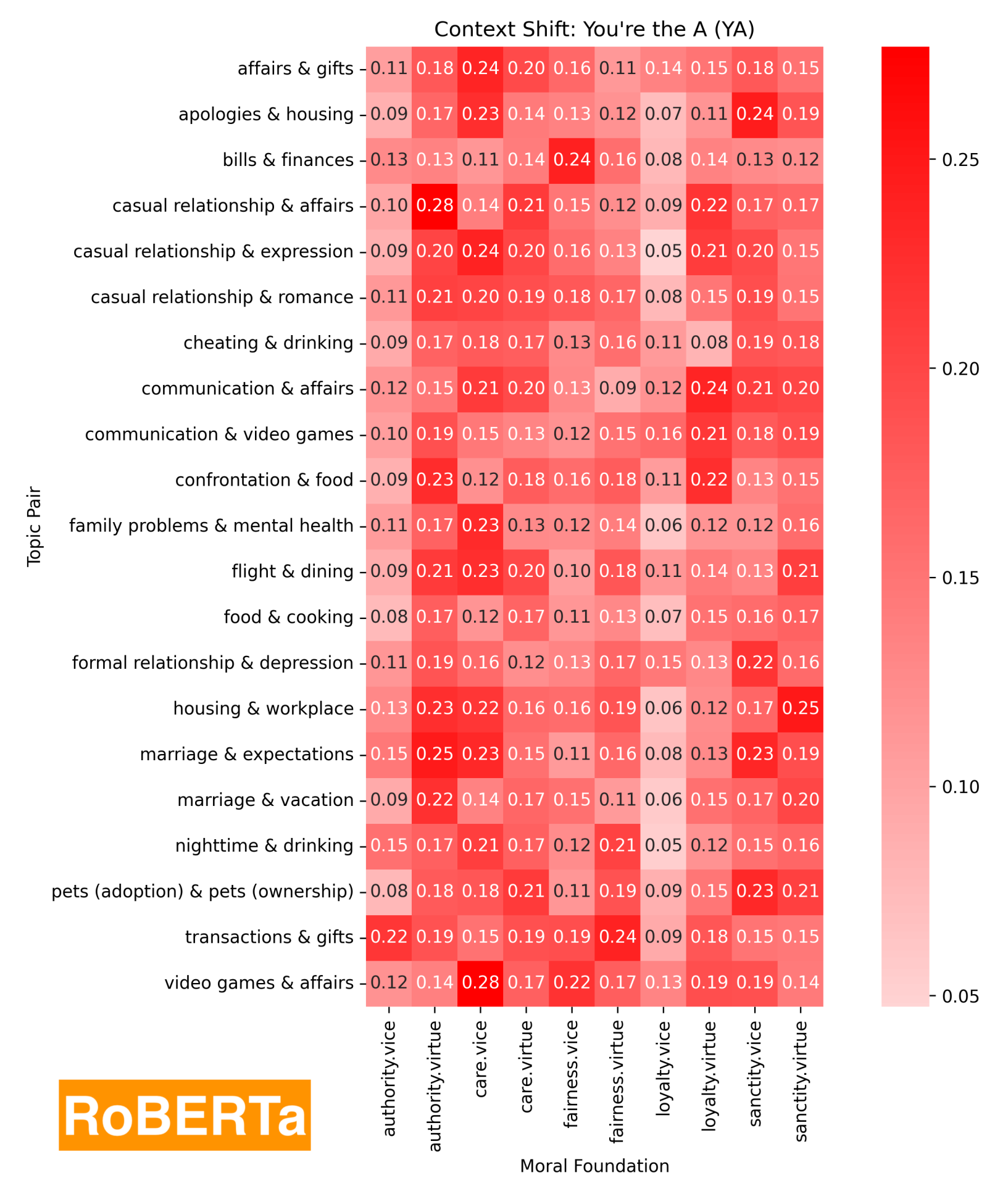
Howdy all. I live in a fairly small city/town in central Oklahoma. We have a fairly large transient population due to 1) high **drug** use around here, 2) being on rail and highway corridors near OKC, 3) a relatively poor local economy, and 4) lots of reservations nearby (four actually within or right next to town). [...]
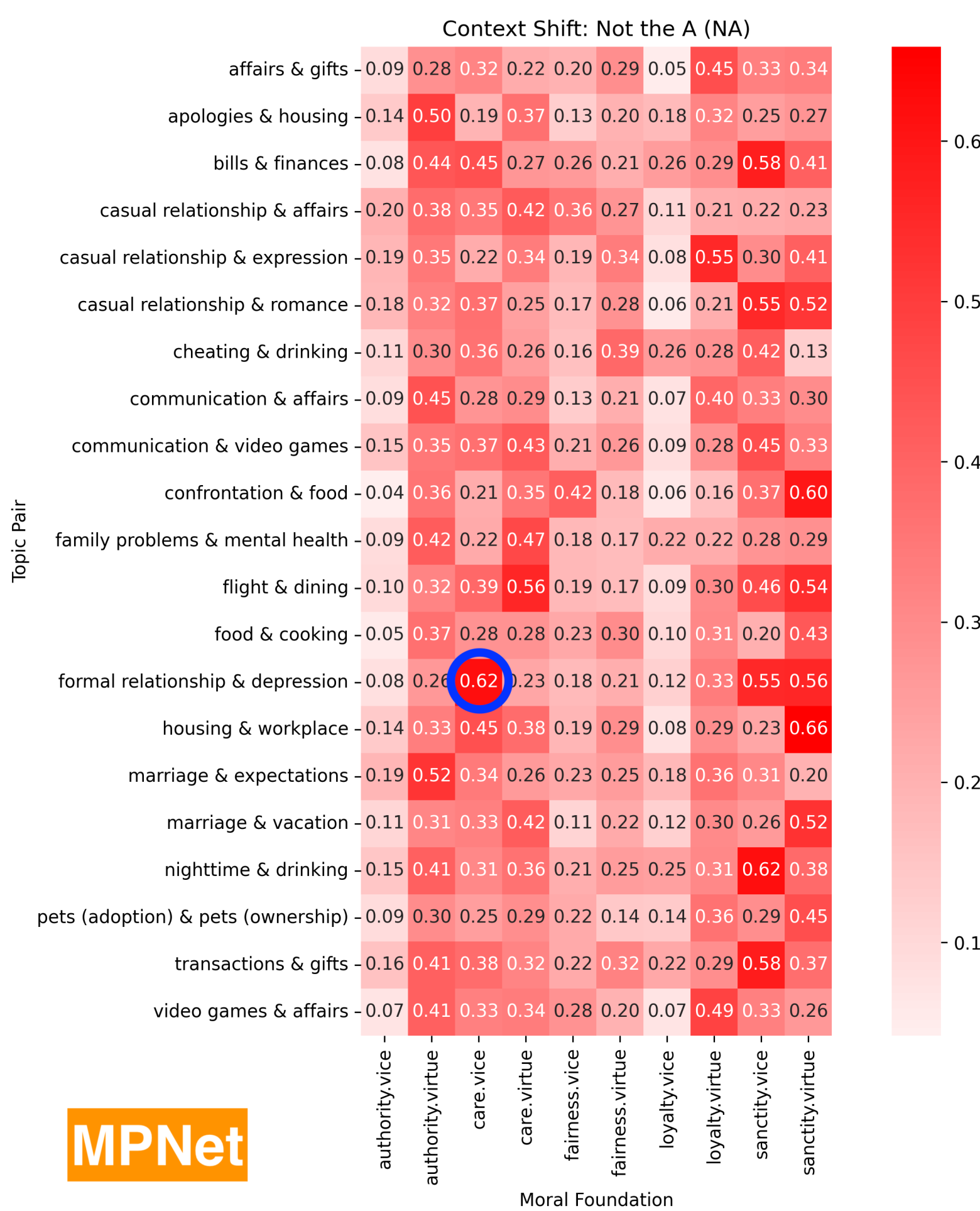
MFD RESULTS

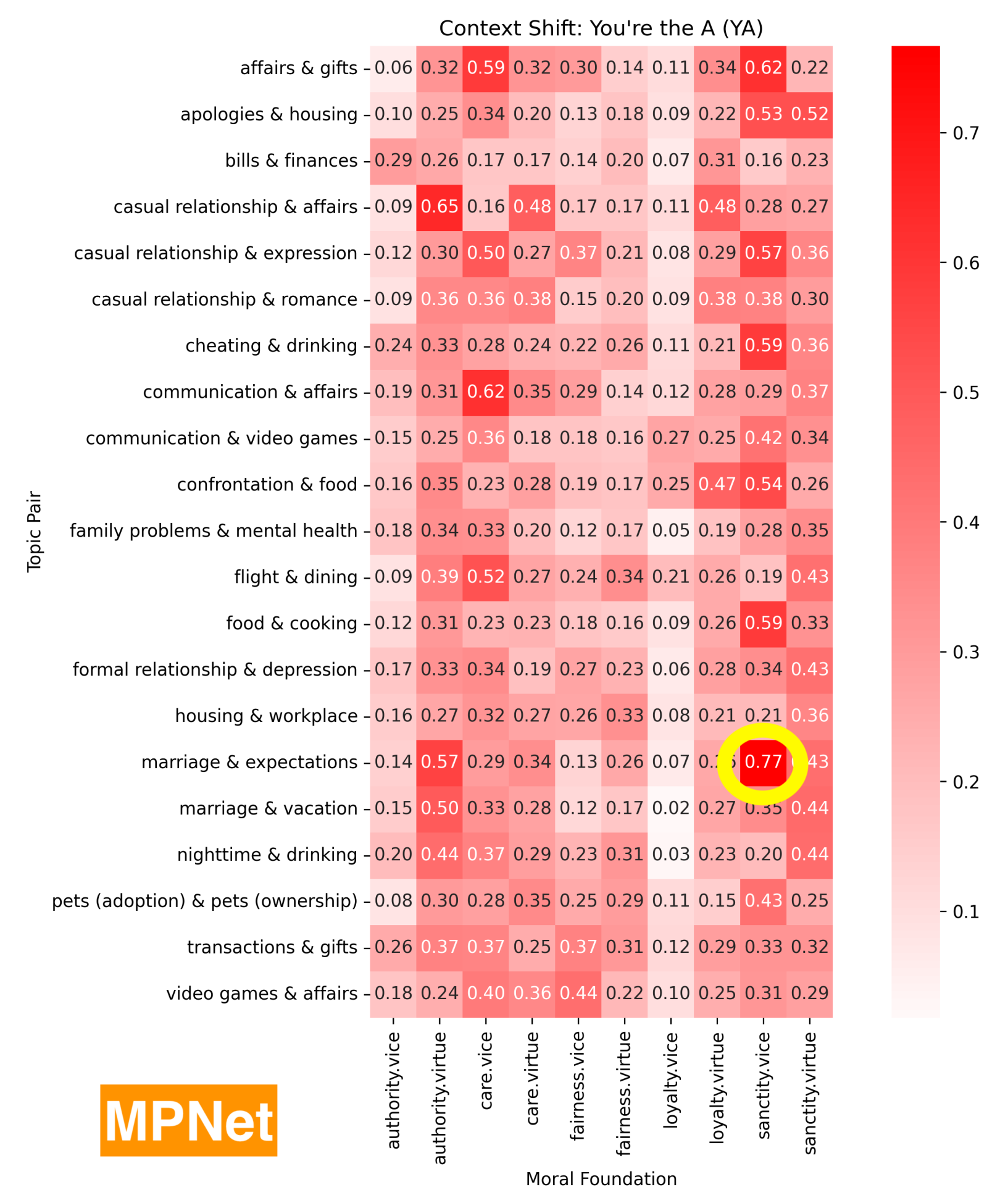Context Shift (Difference) Heatmaps (1 − Cosine Similarity)

Max difference per foundation was plotted

REFERENCE

[1] Tuan Dung Nguyen, Georgiana Lyall, Alasdair Tran, Minjeong Shin, Nicholas George Carroll, Colin Klein, and Lexing Xie. 2022. **Mapping Topics in 100,000 Real-Life Moral Dilemmas.** Proceedings of the International AAAI Con- ference on Web and Social Media 16, 1 (2022), 699–710.