Cornell Natural-Experiment Tweet Pairs v1.0 (released April 2014)

Distributed together with:

The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter
Chenhao Tan, Lillian Lee, Bo Pang
Proceedings of ACL, 2014

The paper, data, demos, and associated materials can be found at:
http://chenhaot.com/pages/wording-for-propagation.html

If you use this data, please cite:
```
@inproceedings{tan+lee+pang:14,
  author = {Chenhao Tan and Lillian Lee and Bo Pang},
  title = {The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter},
  year = {2014},
  booktitle = {Proceedings of ACL}
}
```

Files description:

Data format: each line contains tweet ids and retweet counts for tweet t1 and tweet t2 respectively (t1,n1 t2,n2).
Both tweets contain the same url and were written by the same author.
Example tweet pair, formatted for readability

t1: I know at some point you've have been saved from hunger by our rolling
    food trucks friends. Let's help support them! http://t.co/zg9jwA5j
n1: 2 retweets

t2: Food trucks are the epitome of small independently owned LOCAL businesses!
    Help keep them going! Sign the petition [same URL]
n2: 13 retweets

To retrieve the tweet contents and retweet counts, please use the Twitter API; we are not allowed to distribute a corpus of tweet texts.
Some tweet contents may no longer be available because their authors changed permissions or deleted them, and retweet counts can change from what they were at the time we gathered the data.

The following files contain the data described in much more detail in Section 3 of the paper.

* meaningful_change_pairs.txt:
  11,404 topic- and author-controlled pairs comprising meaningful and significant changes
        (determined automatically by similarity filters; see paper for details)
* holdout_pairs.txt:
  1,770 topic- and author-controlled pairs additional "meaningful change" heldout pairs.
        (For our ACL14 paper, we used this data just once, right before submission, so it was truly unseen data for us.)
* all_pairs.txt:

all 2.4M topic- and author-controlled pairs, including those whose contents are identical up to spacing

Please email any questions to: chenhao@chenhaot.com