Mapping Psychological Sciences through PubMed Text Mining

Muxuan Lyu[1], Yihan Zhang[2], Kyoung Whan Choe[3], & Marc G. Berman[3]

[1]MAPSS, [2]The College, [3]Department of Psychology at The University of Chicago

ENL
Environmental Neuroscience Lab

Introduction

- Literature reviews play an important role in summarizing past research and predicting future research directions
- We applied text mining techniques to construct a high-level overview of current research trends in psychology

Methods

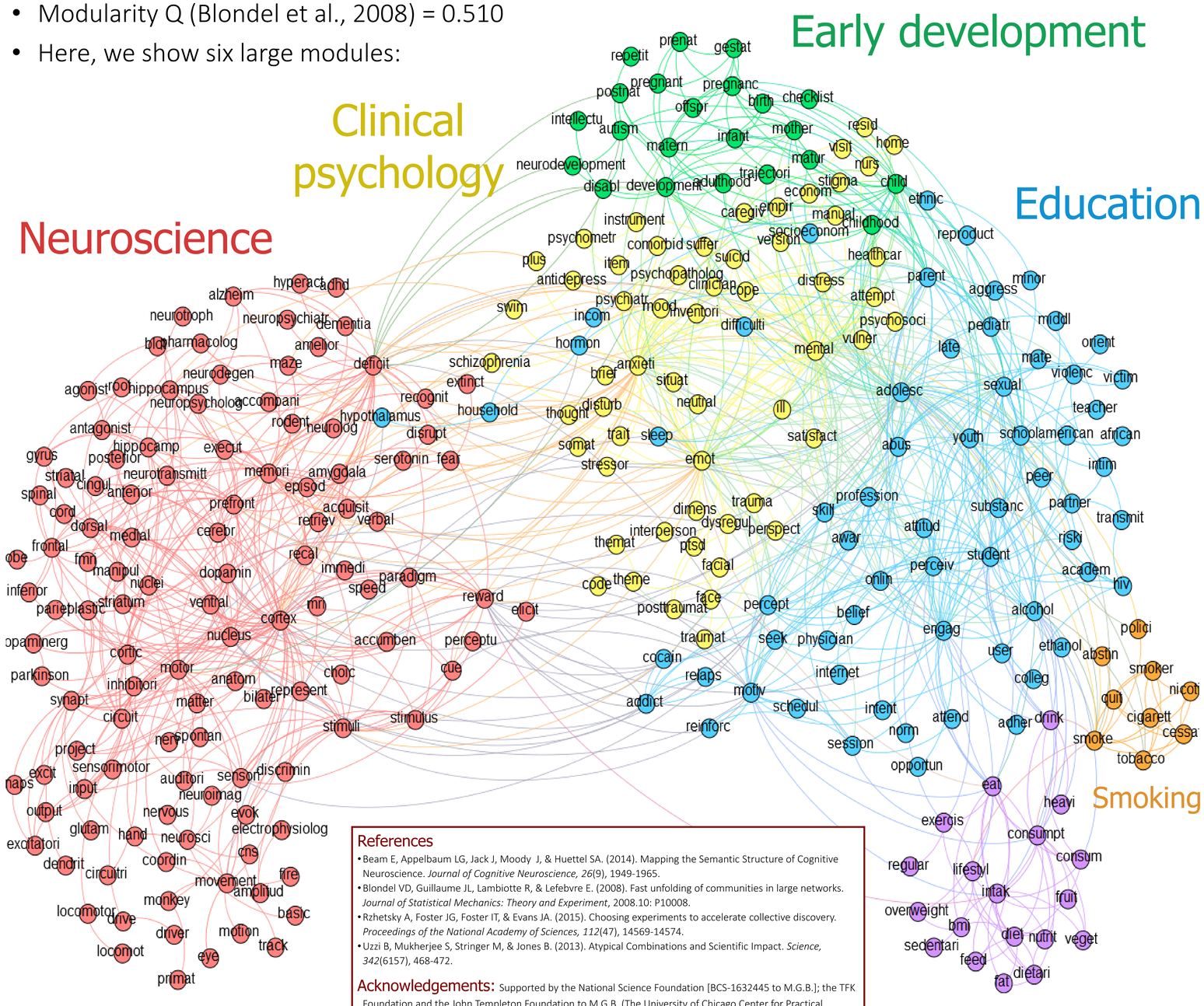
- Tokenized about one million abstracts from PubMed update files, in which 87% were dated from 2016 to 2018, using the snowball stemmer in the python nltk package
- Identified 111,571 abstracts of psychological science (PS) disciplines that contain at least one of the three key tokens: 'behavior', 'psycholog', and 'neurosci'
- Built a dictionary of 847 PS tokens by removing 1,000 top-frequent tokens in the non-PS abstracts from 2,000 top-frequent tokens in the PS abstracts and by visual inspection
- Built an adjacency matrix by counting the number of abstracts in which each pair of tokens co-occurred
- Used Gephi (https://gephi.org/) to visualize the network and ran a modularity analysis. For better visualization, we filtered the edge weight and edge degree between 0.7 to 0.9 and from 1 to 65, respectively

Future Directions

- Use all 30M PubMed abstracts to analyze 1M BS abstracts
- Expand the dictionary to exhaustively cover all PS-related concepts. This may include using n-grams
- Perform time-evolving network analyses (Rzhetsky et al., 2015) to examine the growth of BS knowledge and how PS is different from other disciplines (e.g., hard sciences)
- Add interactive information visualization to help exploring the network and generating promising hypotheses

Results

- The current network has 289 concepts and 1004 edges
- Modularity Q (Blondel et al., 2008) = 0.510
- Here, we show six large modules:



References

- Beam E, Appelbaum LG, Jack J, Moody J, & Huettel SA. (2014). Mapping the Semantic Structure of Cognitive Neuroscience. *Journal of Cognitive Neuroscience, 26*(9), 1949-1965.
- Blondel VD, Guillaume JL, Lambiotte R, & Lefebvre E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008.10: P10008.
- Rzhetsky A, Foster JG, Foster IT, & Evans JA. (2015). Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences, 112*(47), 14569-14574.
- Uzzi B, Mukherjee S, Stringer M, & Jones B. (2013). Atypical Combinations and Scientific Impact. *Science, 342*(6157), 468-472.

Acknowledgements: Supported by the National Science Foundation [BCS-1632445 to M.G.B.]; the TFK Foundation and the John Templeton Foundation to M.G.B. (The University of Chicago Center for Practical Wisdom and the Virtue, Happiness and Meaning of Life Scholars group)

Contact Muxuan Lyu (mlyu@uchicago.edu) for more information.

Mapping Psychological Sciences through PubMed Text Mining

Muxuan Lyu[1], Yihan Zhang[2], Kyoung Whan Choe[3], & Marc G. Berman[3]

[1]MAPSS, [2]The College, [3]Department of Psychology at The University of Chicago

Environmental Neuroscience Lab

Introduction

- Literature reviews play an important role in summarizing past research and predicting future research directions
- We applied text mining techniques to construct a high-level overview of current research trends in psychology

Methods

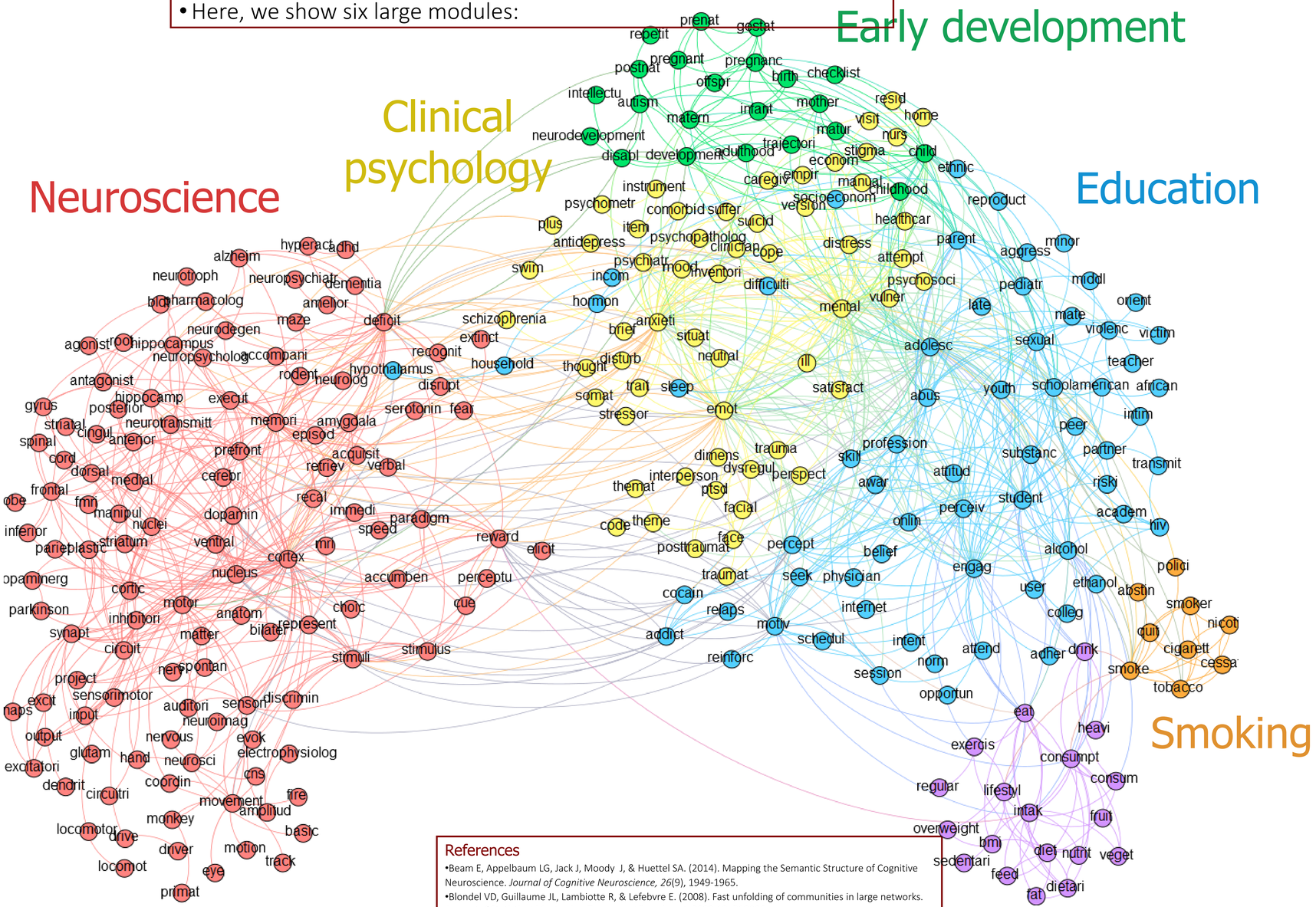
- Tokenized about one million abstracts from PubMed update files, in which 87% were dated from 2016 to 2018, using the snowball stemmer in the python nltk package
- Identified 111,571 abstracts of psychological science (PS) disciplines that contain at least one of the three key tokens: 'behavior', 'psycholog', and 'neurosci'
- Built a dictionary of 847 PS tokens by removing 1,000 top-frequent tokens in the non-PS abstracts from 2,000 top-frequent tokens in the PS abstracts and by visual inspection
- Built an adjacency matrix by counting the number of abstracts in which each pair of tokens co-occurred
- Used Gephi (https://gephi.org/) to visualize the network and ran a modularity analysis. For better visualization, we filtered the edge weight and edge degree between 0.7 to 0.9 and from 1 to 65, respectively

Future Directions

- Use all 30M PubMed abstracts to analyze 1M BS abstracts
- Expand the dictionary to exhaustively cover all PS-related concepts. This may include using n-grams
- Perform time-evolving network analyses (Rzhetsky et al., 2015) to examine the growth of BS knowledge and how PS is different from other disciplines (e.g., hard sciences)
- Add interactive information visualization to help exploring the network and generating promising hypotheses

Results

- The current network has 289 concepts and 1004 edges
- Modularity Q (Blondel et al., 2008) = 0.510
- Here, we show six large modules:



References

•Beam E, Appelbaum LG, Jack J, Moody J, & Huettel SA. (2014). Mapping the Semantic Structure of Cognitive Neuroscience. *Journal of Cognitive Neuroscience, 26*(9), 1949-1965.

•Blondel VD, Guillaume JL, Lambiotte R, & Lefebvre E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008.10: P10008.

•Rzhetsky A, Foster JG, Foster IT, & Evans JA. (2015). Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences, 112*(47), 14569-14574.

•Uzzi B, Mukherjee S, Stringer M, & Jones B. (2013). Atypical Combinations and Scientific Impact. *Science, 342*(6157), 468-472.

Acknowledgements: Supported by the National Science Foundation [BCS-1632445 to M.G.B.]; the TFK Foundation and the John Templeton Foundation to M.G.B. (The University of Chicago Center for Practical Wisdom and the Virtue, Happiness and Meaning of Life Scholars group)

Contact Muxuan Lyu (mlyu@uchicago.edu) for more information.

Introduction

- Literature reviews play an important role in summarizing past research and predicting future research directions
- We applied text mining techniques to construct a high-level overview of current research trends in psychology

Methods

- Tokenized about one million abstracts from PubMed update files, in which 87% were dated from 2016 to 2018, using the snowball stemmer in the python nltk package
- Identified 111,571 abstracts of psychological science (PS) disciplines that contain at least one of the three key tokens: 'behavior', 'psycholog', and 'neurosci'
- Built a dictionary of 847 PS tokens by removing 1,000 top-frequent tokens in the non-PS abstracts from 2,000 top-frequent tokens in the PS abstracts and by visual inspection
- Built an adjacency matrix by counting the number of abstracts in which each pair of tokens co-occurred
- Used Gephi (https://gephi.org/) to visualize the network and ran a modularity analysis. For better visualization, we filtered the edge weight and edge degree between 0.7 to 0.9 and from 1 to 65, respectively

Results

- The current network has 289 concepts and 1004 edges
- Modularity Q (Blondel et al., 2008) = 0.510
- Here, we show six large modules:

Future Directions

- Use all 30M PubMed abstracts to analyze 1M BS abstracts
- Expand the dictionary to exhaustively cover all PS-related concepts. This may include using n-grams
- Perform time-evolving network analyses (Rzhetsky et al., 2015) to examine the growth of BS knowledge and how PS is different from other disciplines (e.g., hard sciences)
- Add interactive information visualization to help exploring the network and generating promising hypotheses