Supplemental Material

(Online via Figshare)

Supplemental Material A: *Additional Details of Latent Dirichlet Allocation*

Latent Dirichlet Allocation (LDA) is a resampling method (pioneered by J. K. Pritchard , M. Stephens , and P. Donnelly in 2000; [1]) for machine learning [2], for application in population genetics. In this report, LDA is used to uncover the sentiment of a corpus of a large number of publications (~10,000 in the current report) to review topics that are discussed tandemly in the literature. From this, albeit non-directional, we can infer relationships between topics, observe how prominent these topics have been over time, observe how relationships have strengthened or waned over time, and discuss the current consensus of the temperature mediated autodigestive response, as well as the broader directions of future research.

The LDA model is "unsupervised" because the statistical inference techniques solely rely on the words contained within a scientific article, free from interpretation prior to the identification of topics. The emergence of a topic "speaks" for itself. Such unbiased inference is useful in the establishment of an emerging area of research by revealing new connections and topical relationships. A naive machine independently synthesizes topics and determines connections between those topics without bias or foreknowledge. "Subjects" are established that are not necessarily limited by traditional disciplines. The subjects are without initial human interpretation and may be as broad as an intersection of many disciplines, or as narrow as an investigation into the relationship between single drugs, conditions, and physiological responses.

A LDA describes a distribution of where a multinomial distribution can land. To sample large multinomial distributions, Markov Chain Monte Carlo (MCMC) algorithms, such as Gibbs Sampling, are employed. Using Gibbs Sampling, topic assignments are resampled iteratively to determine hidden topics that will come to be realized. In doing so, the task of inference is simplified. The topic assignment conditioned on the given data becomes a distribution of our latent variables - such as a document's distribution of topics - conditioned on the Dirichlet parameters. Upon simplifying the Gibbs Equation, we can see that the distribution of our latent variables simplifies to the product of how much a document likes a topic and how much a topic likes a word.

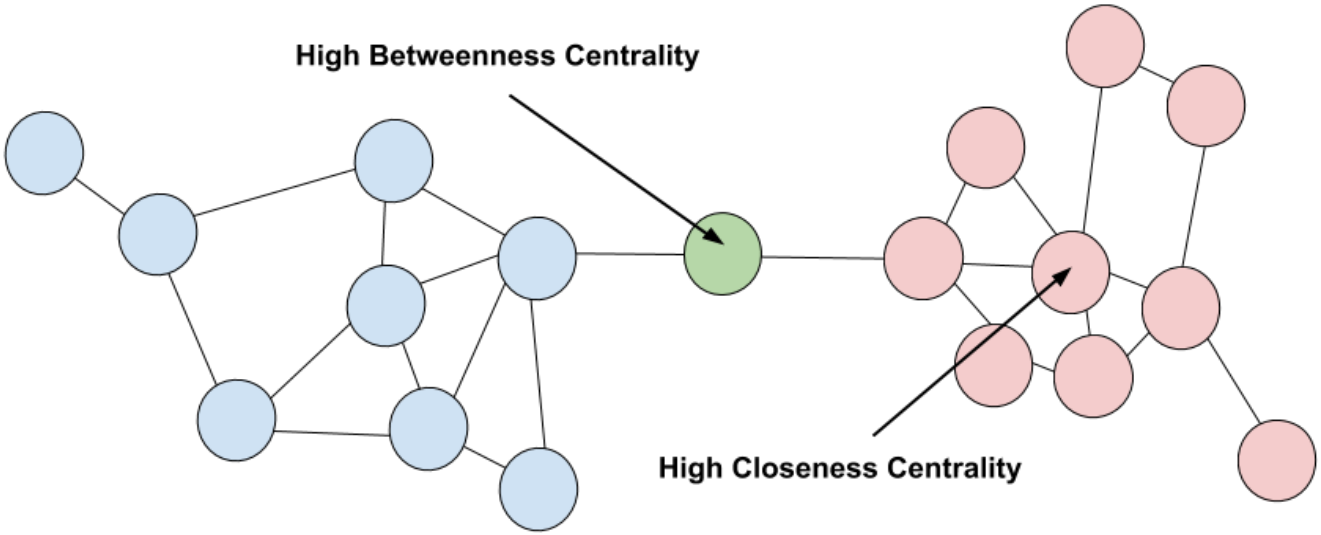
Essentially, LDA seeks to describe the probability a selected topic is discussed in a document, as well as the probability selected words in that document describe a topic. A document may discuss n topics, each to a varying extent that sums to 100%. These topics are unsupervised and are comprised of certain words in the document that are most associated with each other. As an example, a document describing intestine (pet training) and shock (care) may cover epithelium topics (dog topics), permeability topics (cat topics), digestive topics (bird topics), and pancreas (fish) topics to the extent of 35%, 20%, 30%, and 15% respectively. These topic names would be inferred

by a scientist because the resampling algorithm is unsupervised. The words agglomerated by the algorithm under the topic of ‘epithelium’ may contain words like ‘tract’, ‘stomach’, ‘secrete’, ‘digest’ etc. Each of these words also has a score detailing the extent to which they belong to that topic. For instance, ‘activate’ may only weakly be epithelium-related because permeability also uses activate, whereas ‘gastrointestinal’ strongly relates to the topic of pancreas. From these data, we may assume the document primarily discusses epithelium and permeability, as opposed to pancreas. Analyzing many documents like these, we may then assume the extent to which epithelium and permeability are discussed tandemly.

References

1. Pritchard, J.K., M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data.* Genetics, 2000. **155**(2): p. 945-959.
2. Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent Dirichlet allocation.* J. Mach. Learn. Res., 2003. **3**: p. 993–1022.

Supplemental Material B: ***Additional Figures***



Supplemental Figure S.1. A schematic description of measure of centrality for a network. High betweenness centrality can describe important links that relate disparate subjects. High closeness centrality can describe topics that are highly correlated with many other topics and frequently discussed within a subject or discipline. Modularity class is an unsupervised description of communities within a network based on the connectivity and network diameter of certain topics.

Figure S.2

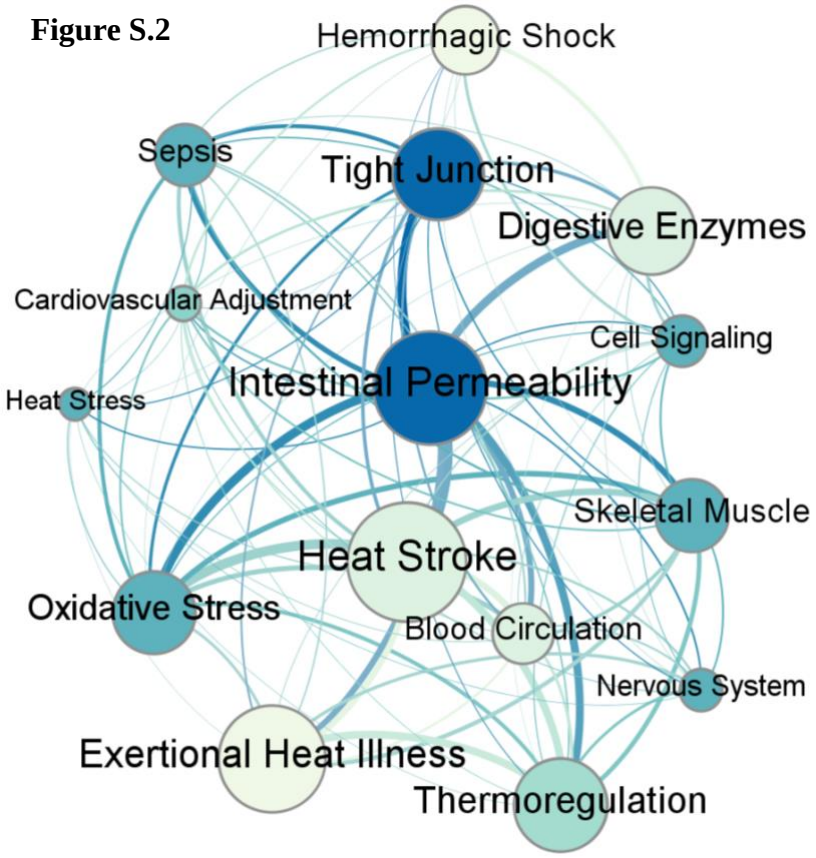
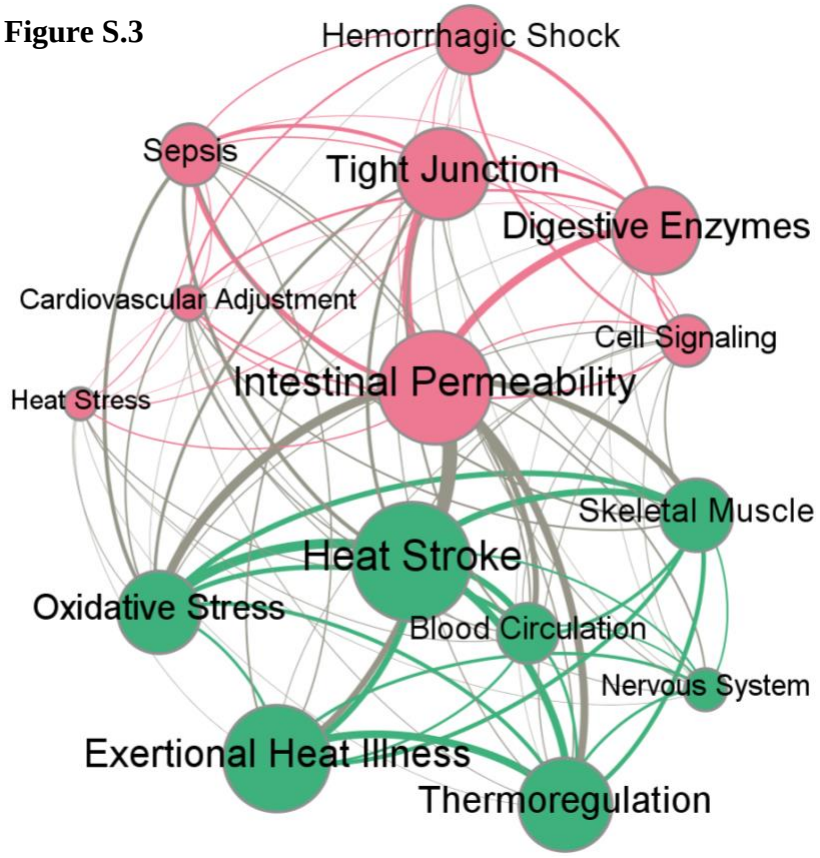
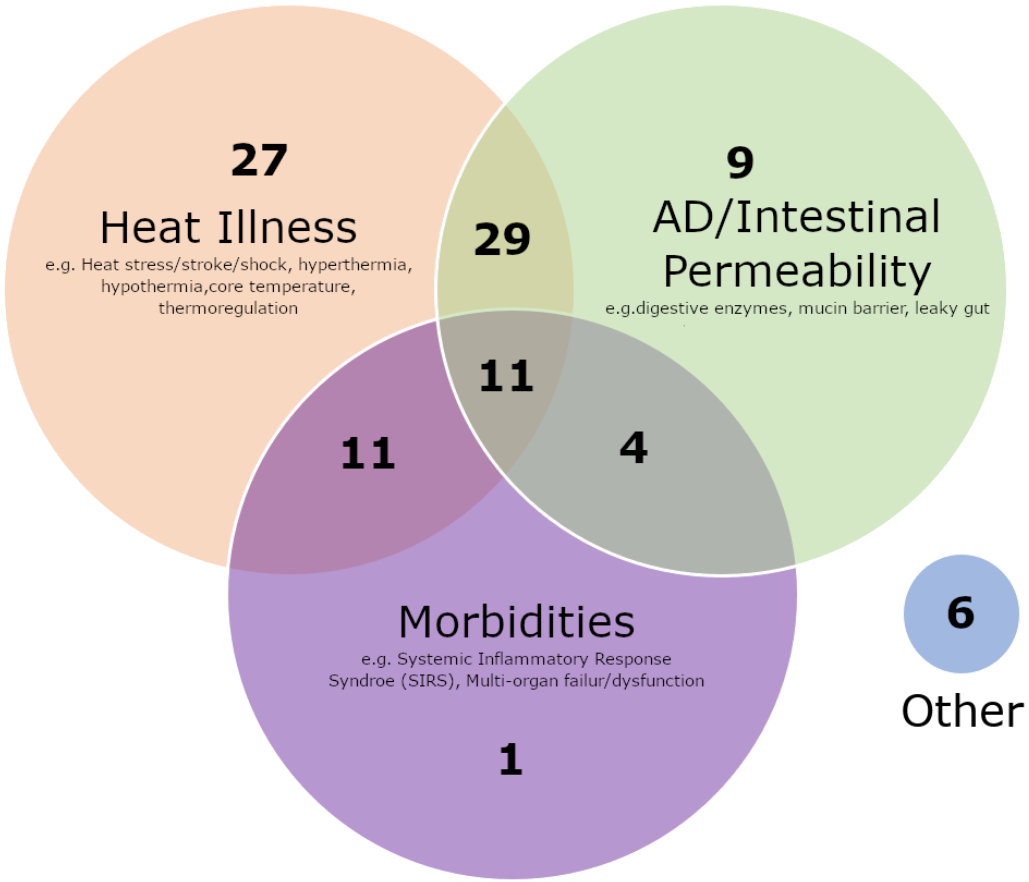


Figure S.3



Supplemental Figures S.2 and S.3: *LDA Relationships of Handpicked Papers* displays the topic distributions uncovered from the papers reviewed in the sections below. **(S.2)** Network map with each node representing a topic. Size of nodes are determined by how often that topic appears in a paper. Edges connecting two nodes are scaled in size by how often the two topics appear in the same report paper. Node color is scaled by betweenness centrality, with darker nodes representing nodes that connect the most topics in any given report paper within the top 10 % percent of TF-IDF filtered papers. **(S.3)** The same network map with automatic modularity class detection. Three categories of topics were uncovered by the network alone, without human intervention or semantic supervision.



Supplemental Figure S.4. Venn diagram of references cited in section 4: Comprehensive review of heat stress mediated autodigestion morbidities. Keyword searches produce results that are not necessarily representative of the global scientific consensus. The 99 references selected for use in section 4 of this paper were chosen based on their published keywords and further analyzed. While this body of literature heavily discusses autodigestion concepts and intestinal permeability, there is a predominance of papers with keywords that fall exclusively in the broader class of heat illness. Co-morbidity keywords, such as inflammation and organ dysfunction, are almost exclusively discussed in tandem with heat stress and autodigestion/intestinal permeability. The intersection of autodigestion/intestinal permeability and heat illness shown here seems very apparent, but may not be reflective of a larger body of literature that focuses on these topics. Furthermore, six papers did not have keywords that would be included in any of the classes uncovered by the LDA analysis, yet provided cogent material for this review. Therefore, a simple keyword search may miss some critical information. This demonstrates that the intersection of subject matter in the literature is highly intricate and broader incorporation and analysis of literature, coupled with human semantic interpretation and insight, may be desirable in uncovering the most critical pieces of information. For these reasons, as well as those described in the main body of this paper, both a high-throughput topic model and traditional manual review (99 studies shown in the Venn Diagram above) were conducted in this manuscript.