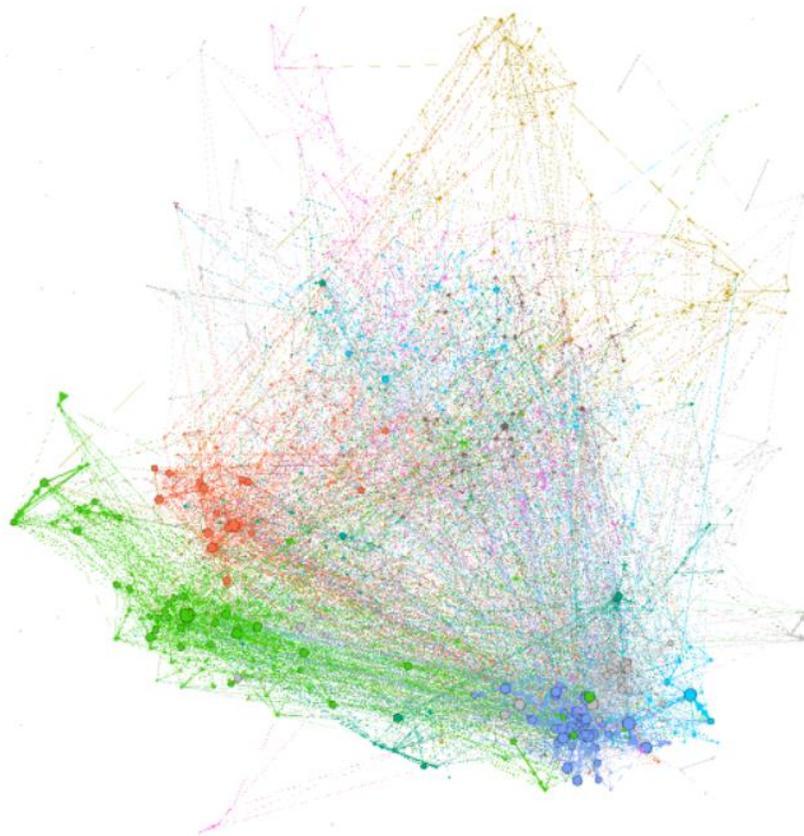


Exploring a Networks of References on the Bible



Sean Letendre
Georgetown University
December 20, 201

I. Introduction

Most Bibles are published with footnotes at the end of the page. In these notes there are often references to other verses or whole chapters in the Bible. These can occur for various reasons from the author quoting older scripture directly to the publishers just wanting to draw attention to thematic similarities across the Bible. Of course these references are not standardized and vary widely between publishers. Nevertheless one could imagine that there exists a network consisting of chapters in the Bible as nodes and these references as edges. The nodes consist of Chapters and not verses because, even though many references point from one verse directly to another, many exist only between chapters as a whole. Secondly, the enormous number of verses would lead to a network too sparse to be analyzed effectively.

Here I have created and explored such a network. I made the network from the New American Bible found at http://www.vatican.va/archive/ENG0839/_INDEX.HTM. This choice was mostly arbitrary, but this site had a lot of links in a format such that the right ones were easy to extract. The goal of this paper is to look at various methods common to network analysis and see which ones do tell us something of significance.

At the end, I include a listing of the [most common] order of books in the Bible with the number of chapters.

II. Summary

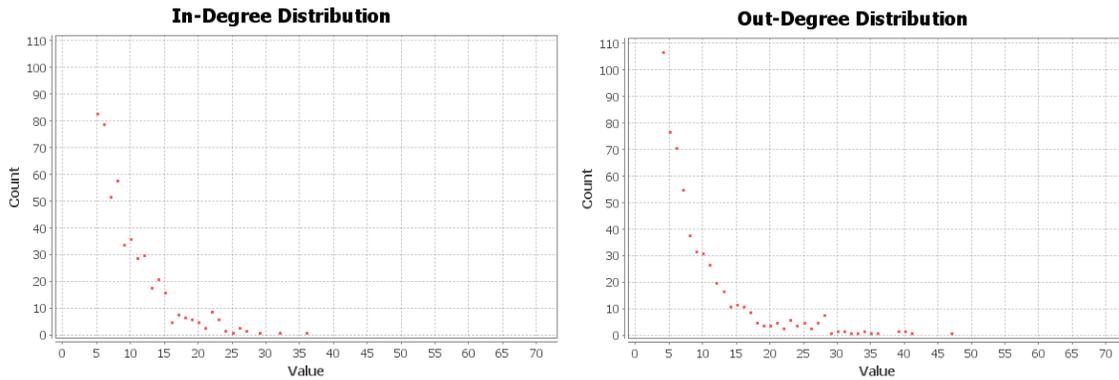
Before going further it will be useful to present a summary of the network, so that the reader will have an idea of what is being worked with. Here is a table giving some of the most common network measures.

Order	Size	Mean In/Out-Degree	Mean Degree	Modularity (Book)	Modularity (Testament)	Reciprocity
1332	8892(6591)	4.95	9.90	0.1782	0.2719	0.3014

Order is the number of nodes present, and size the number of edges. The size in parenthesis is considering multi-edges as just one edge with a weight. For obvious reasons the mean in-degree is equal to the mean out-degree. The mean degree is twice that number. Modularity is a measure of homophily. Modularity will be covered more thoroughly, but in general the greater the number – the more often nodes tend to connect with others of the same type (in this case types are either the Book or the Testament a chapter is in). Reciprocity is the probability an edge is ‘reciprocated’, i.e., the fraction of nodes with source a and target b for which there also exist an edge with source b and target a .

The giant component contains 94.5% of the chapters: 1259 of 1332. Of those nodes not in the giant component, 67 had degree 0 and there were 3 pairs consisting of just one node pointing to the other.

Here are the in- and out-degree distributions:



Strictly speaking, this network is directed: references listed at the bottom of a chapter might not be present at the end of the one it points to (in fact it happens only 30.14% of the time). For many applications the edges on this graph can be thought of as being undirected. This is not an irrational leap: the edges simply say that two chapters have *something* in common which is a symmetric relationship, therefore I will use both in different contexts.

III. Centrality

Here I look at some centrality measures. In a general sense, the goal is to see which chapter of the Bible is the “most important” or “central”. In this case it is obvious neither exactly what this means nor what method should be used. In fact, different measures answer different questions. Here are the top 10 ranked chapters by the most common centrality measures:

	In-Degree	Out-Degree	Degree	Eigenvector Centrality	PageRank
1	Romans 1	Revelation 1	Romans 1	Matthew 10	Genesis 1
2	Matthew 10	Matthew 26	Matthew 26	Matthew 16	Exodus 34
3	Matthew 5	Matthew 27	John 1	Matthew 3	Judges 1
4	1 Corinthians 15	Matthew 21	Mark 14	Matthew 8	Romans 1
5	Mark 14	John 1	Matthew 27	Matthew 17	2 Kings 23
6	Matthew 16	Matthew 8	Matthew 8	Matthew 4	Psalms 89
7	Matthew 3	Romans 1	Matthew 5	Mark 1	Numbers 6
8	Genesis 2	Matthew 19	Revelation 1	Matthew 12	Matthew 27
9	Acts 13	Mark 14	Luke 1	Matthew 13	Exodus 20
10	Luke 6	John 2	Matthew 21	Matthew 5	Joshua 15

Centrality measures are dominated by the New Testament (yellow), especially Matthew (dark yellow), with the exception of PageRank (Old Testament chapters are in red)

The eigenvector centrality required the edges to be counted as undirected. It is defined such that each nodes centrality, x_i , is proportional to that of its neighbors:

$$x_i = \alpha \sum_j A_{ij} + \beta$$

Where A_{ij} is the entry in the i th row and j th column of the adjacency matrix and α & β are arbitrary constants

PageRank, essentially the directed edge version of eigenvector centrality, says that each nodes centrality is proportional to the sum of its neighbors' divided by their out-degrees.

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta$$

Where A_{ij} is the entry in the i th row and j th column of the adjacency matrix, α & β are arbitrary constants, and k_j^{out} is the out-degree of node j

Of course, the in-degree column tells us which chapters are referenced the most. The majority are from the New Testament, but we still have Genesis. Romans 1 ranks the highest because it contains a summary of Paul's theology of sin and salvation, the things he discusses most.

In the out-degree column we see which chapters depend the most on previous writing. It makes sense that Revelation comes in at the highest as it is the last book of the Bible.

With regard to Eigenvector centrality we see something pretty strange: nine chapters from Matthew. This means that that Matthew is frequently referenced by other Gospels, including itself. This can be explained.

PageRank tells a different story. It makes sense that chapters from the Old Testament appear mostly in the PageRank column, because these are rewarded for having few out-edges. These books, conversely chapters will often refer back, and those ones will refer further back. In the PageRank column we see some of the most important passages of the Old Testament. Genesis 1 has the first creation story, Exodus 34 contains the 10 Commandments. 2 Kings 23 has King Josiah reading aloud the contents of the recovered "Book of the Law". In short, page rank tells us which chapters are most important for creating persistent themes.

These rankings are consistent with an interesting structure in the network. There is one 14-core (a k -core is a maximal subgroup of nodes such that each node is connected to at least 14 others – in this case the graph is again taken to be undirected) of size 62 that contains 56 of 68 chapters in the synoptic gospels. It should come as no surprise that the synoptic Gospels form such a core as they essentially three version of the same story.

The presence of the core explains why so many of the Gospel chapters have such high centrality. Additionally these are considered the most important books to Christianity, so they should rank high.

IV. Community Detection

Before exploring community detection, I will look at the other side of the problem relating to groups and talk more about homophily. For networks with nodes of discrete types, homophily measures the level to which nodes tend to associate with others of the same type. In more concrete terms, it is usually measured by *modularity*. The modularity, Q , of a network is given as follows:

$$Q = \frac{1}{2m} \sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j)$$

Where m is the size of the network A_{ij} is the entry in the i th row and j th column of the adjacency matrix, k is the degree of a node, and c is the type of a node. $\delta(x,y)$ evaluates to 1 if $x=y$, and 0 otherwise.

This formula can be used on networks with weighted entries. Entries higher than 1 in the adjacency matrix will simulate there actually being more edges. When c was taken from the set consisting of book names, modularity was calculated to be $Q=0.1782$. When c took values “Old Testament” or “New Testament”, the modularity was calculated to be $Q=0.2719$. This number being less than 0.5 means that the majority of references were inter-testamental. These numbers being relatively low implies that there is something other than chapters and Testament that underlying the structure of our network.

The opposite side of the coin is to do *community detection*: dividing all nodes into any number of categories, c_i , in whichever way that modularity. With large networks, it is too computationally complex to actually find this grouping exactly. Instead it is necessary to use some heuristic algorithm, the most effective one at the moment is the Louvain Modularity algorithm summarized below:

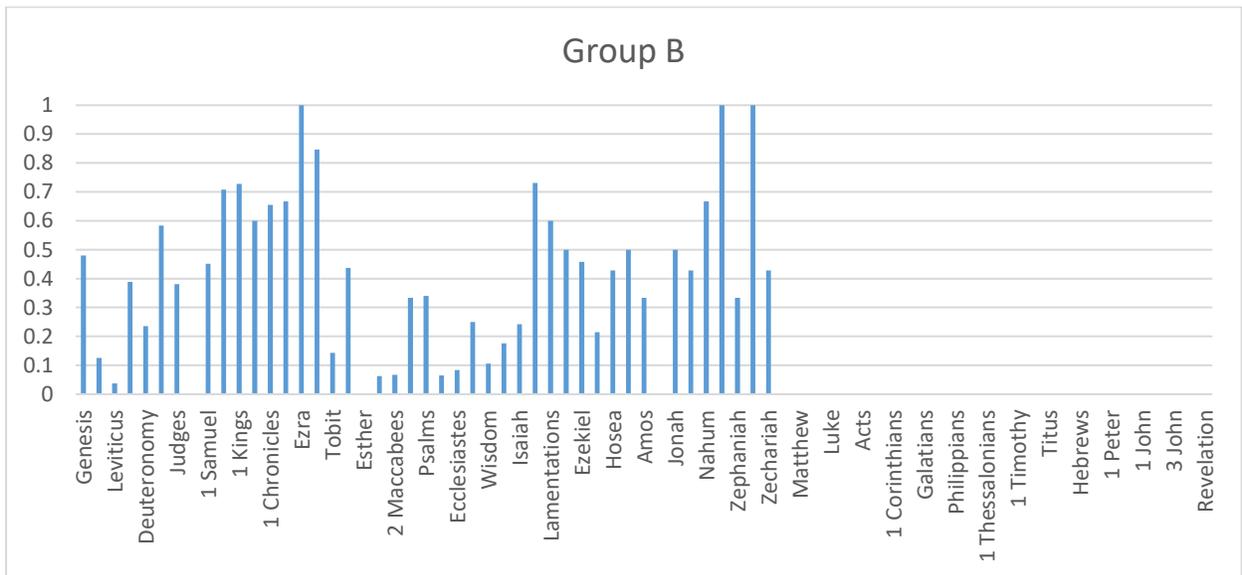
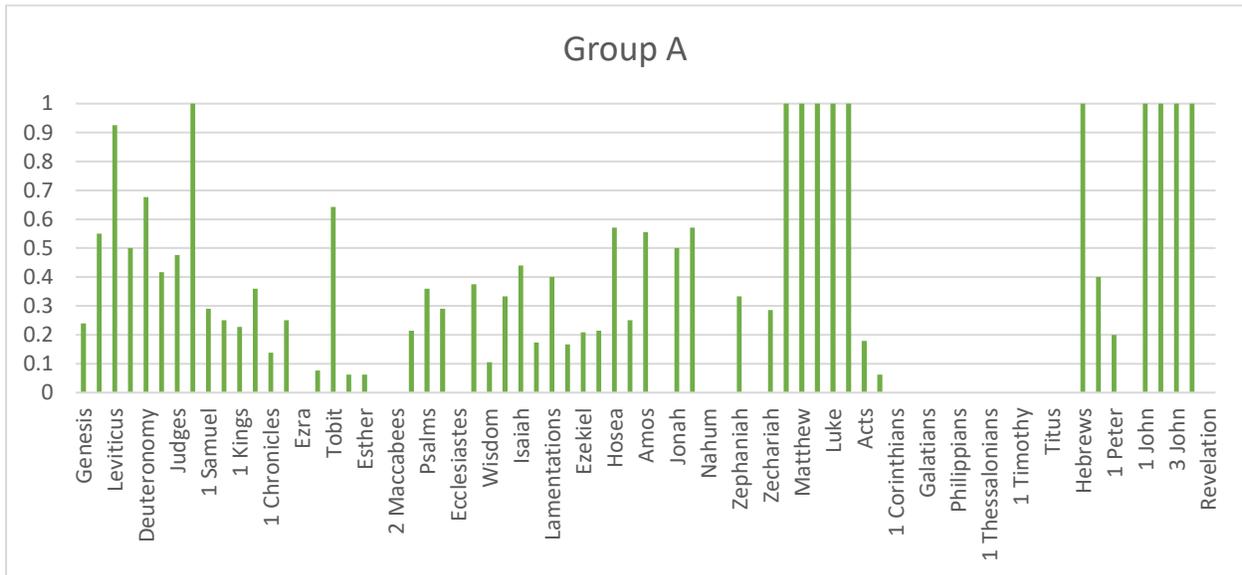
Until modularity stops increasing, repeat the following:

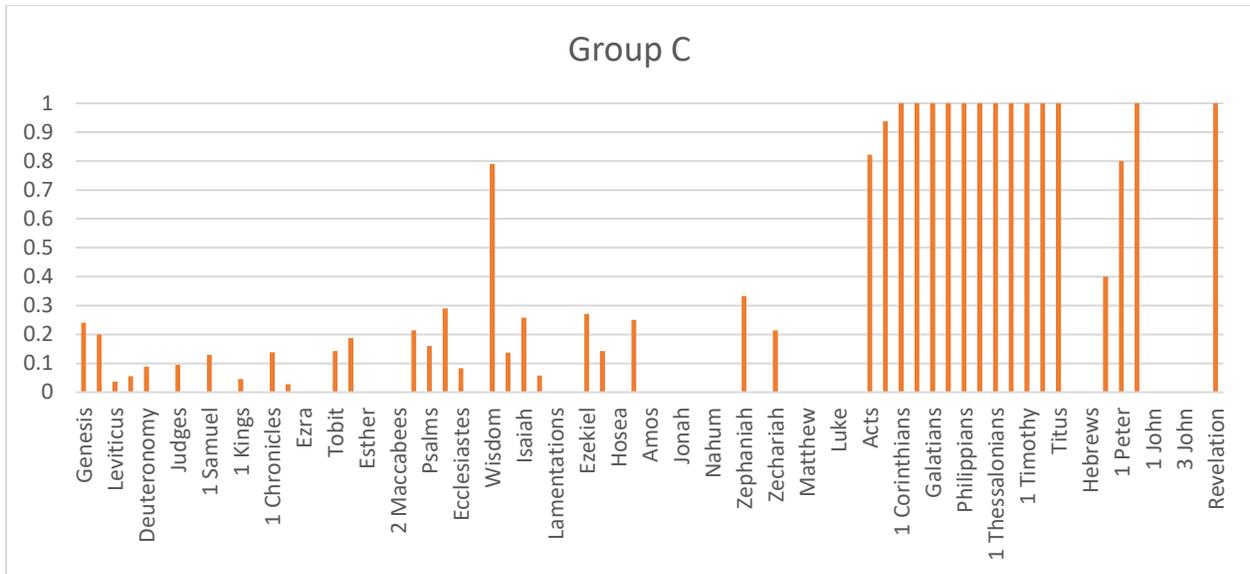
- **Call each node it's own community**
- **For each node n (in random order):**
 - **For each of n 's neighbors i :**
 - **Calculate the increase, P , in the networks modularity that would occur if i joined n 's community**
 - **Add the neighbor corresponding to the largest value P to n 's community**
- **Combine all nodes in a community into one node. All internal edges become one weighted self-loop, and all external edges to a given community become one weighted edge to the corresponding new node**
- **Calculate the modularity of the network as a whole, that is not considering communities as one node**

Where *modularity* is defined exactly as it was above.

The software *Gephi* provides an implementation of the Louvain Modularity algorithm which I used. After running this algorithm, three significant communities were found and many others of trivial size. This is how I will refer to the main groups of significant size: Group A containing 35.51% of chapters, Group B containing 30.56% of chapters, and Group C containing 21.55% of chapters.

The question is, do these groupings represent any real similarities. To figure this out without reading each chapter and contemplating common themes, I did a sort of “spectroscopy”. I plotted the number of chapters each group by book. It turns out that by dividing each count by the total number of chapters in that book, the patterns are easier to spot. Along the horizontal, axis are the names of only some books. For the complete listing, in proper order, consult the list of books at the end of the paper.





These communities do in fact seem to be decipherable:

- Group A is about Jesus. It contains 100% of all four Gospels, Malachi (a messianic prophecy), Hebrews (an essay on the nature of Christ), the very short books of 1-3 John and Jude (still about Jesus), and Ruth (Ruth is only four chapters long, so this is not saying much).
- Group B is about the nation of Israel. First of all, this group contains nothing from the New Testament. Secondly it emphasizes the history of Israel. Golden Age contains all of the books Ezra, Haggai, and contains 85% of Nehemiah – books about the return to Israel after exile. It contains all of Habakkuk as well, which is only three chapters long.
- Group C is about Paul, the author of about half of the New Testament. It contains all of his writings and other letters in the same vein of literature. Group C also has 79% of the book of Wisdom. This is not an anomaly, it is often referenced to explain Paul's background in Hellenistic-Jewish philosophy (Wisdom is known for being influenced by Greek philosophy).

These communities are quite large - therefore their existence is proof that this algorithm works, but it does not necessarily provide us with any new information. I was expecting to find a large number of small networks defined by a more focused theme, such as marriage or baptism as examples.

V. Conclusion

The information gained from standard network analysis techniques was indeed acceptable. In most cases results were consistent with what we know about the Bible. In general, it was shown that literary networks are not beyond the scope of network analysis and that in the future similar networks could be analyzed.

Additionally, what was concluded by the research is that Pauline Letters, especially Romans, and the Gospels, especially Matthew, are the most important books to our [or at least this publisher's] understanding of religion.

Resources

- I used equations and measures found in the book:
Networks by M.E.J. Newman
published by Oxford University Press Inc., New York City
- The network data were collected from the website:
http://www.vatican.va/archive/ENG0839/_INDEX.HTM
with the Java language with the *Jsoup* package :
jsoup HTML parser © 2009 - 2016 Jonathan Hedley
- Other measures and all network pictures were made with the software *Gephi* :
Gephi.org © All Rights Reserved 2008-2016
- I summarized the algorithm presented in the paper:
Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre,
Fast unfolding of communities in large networks, in *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10), P1000

The Order of Books in the NASB

The Old Testament

Pentateuch

- Genesis (50)
- Exodus (40)
- Leviticus (27)
- Numbers (36)
- Deuteronomy (34)

Historical Books

- Joshua (24)
- Judges (21)
- Ruth (4)
- 1 Samuel (31)
- 2 Samuel (24)
- 1 Kings (22)
- 2 Kings (25)
- 1 Chronicles (29)
- 2 Chronicles (36)
- Ezra (10)
- Nehemiah (13)
- Tobit (14)
- Judith (16)
- Esther (16)
- 1 Maccabees (16)
- 2 Maccabees (15)

Wisdom Books

- Job (42)
- Psalms (150)
- Proverbs (31)
- Ecclesiastes (12)
- Song of Songs (8)
- Wisdom (19)
- Ben Sirach (51)

Prophetic Books

- Isaiah (66)
- Jeremiah (52)
- Lamentations (5)
- Baruch (6)
- Ezekiel (48)
- Daniel (14)
- Hosea (14)
- Joel (4)
- Amos (9)
- Obadiah* (1)
- Jonah (4)
- Micah (7)
- Nahum (3)
- Habakkuk (3)
- Zephaniah (3)
- Haggai (2)
- Zechariah (14)
- Malachi (3)

* The book of Obadiah was, by accident, not included in the network. As it is only one chapter long and no references were made to it, the information presented above should remain unaffected.

The New Testament

Gospels

- Matthew (28)
- Mark (16)
- Luke (24)
- John (21)
- Acts (28)

Letters

- Romans (16)
- 1 Corinthians (16)
- 2 Corinthians (13)
- Galatians (6)
- Ephesians (6)
- Philippians (4)
- Colossians (4)
- 1 Thessalonians (5)
- 2 Thessalonians (3)
- 1 Timothy (6)
- 2 Timothy (4)
- Titus (3)
- Philemon* (1)
- Hebrews (13)

General Letters

- James (5)
- 1 Peter (5)
- 2 Peter (3)
- 1 John (5)
- 2 John (1)
- 3 John (1)
- Jude (1)
- Revelation (22)

* The book of Philemon was, by accident, not included in the network. As it is only one chapter long and no references were made to it, the information presented above should remain unaffected.