Chapter 0

Introduction

Origin of the Project

The most significant finding of my dissertation is that the author who wrote the thirty-six case statements introducing the hypothetical cases that make up the second part of Gratian's *Decretum* is very unlikely to have been the same person as the author who wrote the *dicta* in the first recension of the *Decretum*. The statistical method used to make this determination takes the frequencies of common function words like prepositions and conjunctions in a sample of text as the basis for assigning probable authorship, and will be explained in considerable detail in Chapter 4.

I did not start work on this project thinking that the authorship of the case statements was in any way a research problem. I assumed that by definition the author of the case statements was one and the same person as the author of the first-recension *dicta*. It is therefore worth explaining in some detail how I came to write a PhD dissertation about a completely unexpected finding that I was not looking for in the first place.

I worked in information technology as a system administrator and manager for most of the twenty-three years after I graduated from UC San Diego in 1984 with an undergraduate degree in History. Stanley Chodorow had been the advisor for my undergraduate senior thesis on the role of the cardinals in the thirteenth and fourteenth centuries, and I knew that he had written a book about Gratian's *Decretum*.[1] I was therefore aware of Gratian in a general sort of way, although the only use I made of the *Decretum* in connection with my thesis was to consult Emil Friedberg's 1879 edition for the Latin text of Nicholas II's 1059 decree on papal elections (D.23 c.1).

Chodorow urged me to use computer-aided typesetting for the project, and in this way I acquired a then-unusual skill that led directly to my IT career. In the mid to late 80s I went on to take most of the required courses for the undergraduate Computer Science major at UC San Diego (e.g., Data Structures, Compiler Construction, Operating Systems) although I did not enroll in a degree program. During my professional career, I was never primarily a programmer, but from time to time my job responsibilities did

---

[1] Stanley Chodorow, *Christian Political Theory and Church Politics in the Mid-Twelfth Century; the Ecclesiology of Gratian's Decretum*, Publications of the Center for Medieval and Renaissance Studies, U.C.L.A., 5 (Berkeley: University of California Press, 1972).

include programming projects in C and Perl, and ultimately Java servlet-based web applications.

In October 2003, quite by accident, I became aware of Anders Winroth's *The Making of Gratian's Decretum*.[2] I had done a Google search for Chodorow, looking for his contact information, and found his review of Winroth's book in *The English Historical Review*.[3] From the review I learned that Winroth had identified five twelfth-century manuscripts as a first, and more coherent, recension of the *Decretum*. In addition, I became aware of Winroth's claim that two different authors, Gratian 1 and Gratian 2, were responsible for the first and second recensions. It was clear to me that there had been a revolution in Gratian studies. My wife Carol gave me the book for Christmas 2003 with the inscription "I'm sure you'll gulp this one down within 24 hours." I did. Some years later, Anders thanked her for buying a copy: "I'm sure I did something very useful with the money".

---

[2] Anders Winroth, *The Making of Gratian's Decretum* (Cambridge: Cambridge University Press, 2000).

[3] Stanley Chodorow, "Review of the Making of Gratian's Decretum by Anders Winroth," *The English Historical Review* 118, no. 475 (February 2003): 174–76.

From September 2007 to May 2009, I was a student in the History of Christianity MAR program at Yale Divinity School. Among the courses I took was a one on Latin Paleography that Richard and Mary Rouse of UCLA taught in the Beinecke Rare Book and Manuscript Library in Spring 2009. Although I had a general interest in applying my computing background to my academic work, I do not think I had heard of Digital Humanities as an academic discipline before I graduated from YDS, at least not by that name.

In October 2009, David Ganz (then of King's College, London) suggested that I compare two texts of the *Capitulare Carisiacense* (873) in Beinecke MS 413. At first, I did not think of this as a digital project; it was simply a transcription exercise of the kind the Rouses had taught me to do. But within a month, I had created a custom text-encoding format for my transcriptions and written a prototype textual difference visualizer in Perl to compare them. A January 2010 meeting with Barbara Shailor on the Beinecke 413 project was the occasion for the first use I can find in my own notes of the term *Digital Humanities*.

In August 2010, I started the PhD program in the Medieval and Byzantine Studies Program at The Catholic University of America in Washington, DC. I went to CUA specifically to work with Ken Pennington on Gratian's *Decretum*. Even before moving from New Haven to Washington, I had participated in Winroth's class on law in Medieval Europe at Yale, and once at CUA, I took Pennington's classes on canon and Roman law, and (twice) his sources seminar. From 2010 through 2012, then, I thoroughly immersed myself in the scholarly debates surrounding Gratian and the *Decretum* with considerable intellectual interest but also a certain level of personal discomfort at being unable to reconcile the contradictory positions staked out by Pennington and Winroth.

Pennington and his students Melodie Harris Eichbauer and Atria A. Larson argued that the Sankt Gallen 673 (Sg) manuscript represented, at however many removes, an earlier version of the *Decretum* than Winroth's first recension; that a single author, Gratian, compiled and wrote both the first and second recensions of the *Decretum*; and for an early date, in the 1130s, for the first recension. Winroth and his student John Wei argued that Sg was a relatively uninteresting abbreviation of a first recension manuscript with some second recension interpolations; that two different authors, Gratian 1 and Gratian

2, compiled and wrote the first and second recensions; and for a late date, around 1140, for the first recension.[4]

In a January 2011 advising conversation, Jennifer Davis suggested that, given my professional background, it would be strategically advantageous for the purpose of whatever academic career I might hope to have to position myself as a Digital Humanities specialist. In the summer of 2010, I had taught myself to write Python web applications on the Google App Engine platform (learning Python was incidental to learning GAE, which is what I was really interested in), so in the first half of 2011, I developed Ingobert, a Python/GAE web application to visualize textual differences in Beinecke 413, in connection with an independent study project supervised by Pennington and Davis.[5] Largely on the strength of the Ingobert project, Neil Fraistat of the University of Maryland hired me as a graduate assistant at the Maryland Institute

---

[4] See Melodie H. Eichbauer, "Gratian's Decretum and the Changing Historiographical Landscape," *History Compass* 11, no. 12 (December 2013): 1111–25 for a good recent overview of these debates.

[5] Ingobert was named after the Carolingian scribe of the Bible of San Paolo fuori le Mura. Some scholars have suggested that he was responsible for Beinecke 413; the script is certainly similar to his. The Ingobert project is still under active development: see my GitHub Ingobert2 repository for the source code of the current version of the Python web application ported to the Django platform.

for Technology in the Humanities (MITH) to work as a Scala/Lift programmer on the Active OCR project.[6]

I finished my PhD comprehensive examinations in October 2012 and advanced to candidacy in January 2013. I had not yet made a definite decision to pursue a dissertation project with a Digital Humanities component, but audited Matt Kirschenbaum's graduate Introduction to Digital Humanities course at the University of Maryland in Spring 2013, with the idea that an overview of the field might suggest a potential project.

My first step was to obtain an electronic version of the *Decretum* text. In the mid- to late-1980s, Timothy Reuter and Gabriel Silagi edited the *Wortkonkordanz zum Decretum Gratiani*, a computer-generated concordance in the tradition of Father Roberto Busa's *Index Thomisticus*, for the Monumenta Germaniae Historica (MGH) in Munich.[7] As part of the project, the MGH undertook to scan, correct, and encode in the now-obsolete and

---

[6] NEH ODH Grant number: HD-51568-12

[7] Timothy Reuter and Gabriel Silagi, eds., *Wortkonkordanz Zum Decretum Gratiani*, Monumenta Germaniae Historica. Hilfsmittel 10 (München: Monumenta Germaniae Historica, 1990).

non-tree-structured Oxford Concordance Program (OCP) format the 1879 Friedberg edition of the *Decretum*. In spring 2013, Winroth and Lou Burnard of the Oxford Text Archive (OTA) each provided me with a copy of the Reuter and Silagi e-text. The two copies, however, differed in many places, and I had to go through a process similar to preparing a critical edition to restore the e-text to a state as close as possible to what I though the editors intended. I then began to experiment with writing Python programs that used regular expressions to extract textual features of interest. The fact that the OCP e-text format is not tree-structured the way XML is—textual features have start tags but not end tags—makes it extremely difficult to parse, so this was a slow process.

My initial focus in the first half of 2013 was on the use of David Mimno's MALLET (MAchine Learning for LanguagE Toolkit) to topic model *dicta* and canon texts from the first and second recensions of Gratian's *Decretum* as a way to identify new topics added in the second recension. The model was Pennington's observation that most passages in the *Decretum* dealing with the legal status of Jews, particularly those dealing with

forced conversion, were introduced only in the second recension.[8] My goal was to see whether MALLET could surface more such topics, by topic modeling the first and second parts of the vulgate *Decretum*, topic modeling the first recension, and seeing what topics were left when the first recension topics were subtracted from the vulgate topics. While simple in concept, this proved prohibitively difficult in practice.[9]

---

[8] Kenneth Pennington, "The Law's Violence Against Medieval and Early Modern Jews," *Rivista Internazionale Di Diritto Comune* 23 (2013): 23–44; and Kenneth Pennington, "Gratian and the Jews," *Bulletin of Medieval Canon Law* 31, no. 1 (2014): 111–24.

[9] This project was attractive to Pennington because although the results would be obtained computationally, they could be verified by someone doing a close reading of the text of the *Decretum*. There were three insurmountable barriers to carrying out the project as originally conceived: the time required to prepare the necessary text samples; the difficulty in determining the number of topics to look for (a necessary precondition for unsupervised topic modeling); and the fact that there was no obvious way to subtract topics.

While a stylometric analysis for authorship attribution requires only the *dicta* (*ante*, *post* and *init.*) thought to have been written by Gratian himself, a topic can be present in any text in the *Decretum*, inscriptions and canons as well as rubrics and *dicta*. It took six weeks—twice—just to prepare a proxy text for the first-recension *dicta*. (In late Summer 2015 I discovered quality anomalies in the *dicta* samples I had hand-edited in Fall 2013, so in Fall 2015, I regenerated the *dicta* samples from scratch by rigorously cross-checking all of the hand-edited *dicta* against a data set automatically generated using Python regular expressions until no differences remained between the two sets of samples.) There is about four times as much text by word count in the canons as there is in the *dicta*, so I estimated that it would take just under six person-months to prepare a proxy text for the first-recension canons.

The Latent Dirichlet Allocation (LDA) algorithm that MALLET uses to generate topic models has to be provided with an exact number of topics to look for. In February 2014, I carried out a preliminary experiment to obtain a rough estimate of the number of topics in the *Decretum*, inspired by the metaphor

In July 2013, I was working at MITH, and following the DH 2013 conference at University of Nebraska-Lincoln out of general interest rather than any sense that it might be relevant to my decision regarding a dissertation topic. One presentation in particular caught my attention: "Stylometry and the Complex Authorship in Hildegard of Bingen's Oeuvre" by Mike Kestemont, Sara Moens, and Jeroen Deploige. Their work was later published as a paper, but the conference website had an unusually detailed abstract, and a video was made available as part of the presentation.[10]

The applicability of Kestemont's methodology to the intractable problem of the authorship of the *Decretum* was immediately obvious to me; it seemed to finally offer a way past endless debates based on indirect evidence about whether there had been one

---

of focusing a telescope. I took the second-recension *dicta* and repeatedly ran MALLET on them, looking for values of the number of topics at which Pennington's topic on the legal status of Jews came into focus. **Pennington's topic started appearing somewhere over 200 topics.**

[10] Abstract: Mike Kestemont, Sara Moens, and Jeroen Deploige, "Stylometry and the Complex Authorship in Hildegard of Bingen's Oeuvre," in *Digital Humanities 2013: Conference Abstracts* (Lincoln, NE: University of Nebraska–Lincoln, 2013), 255–58, http://dh2013.unl.edu/abstracts/ab-126.html. Video: Mike Kestemont, "Documentary: "Hildegard of Bingen: Authorship and Stylometry" [HD]," July 18, 2013, https://vimeo.com/70881172. Paper: Mike Kestemont, Sara Moens, and Jeroen Deploige, "Collaborative Authorship in the Twelfth Century: A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux," *Literary and Linguistic Computing* 30, no. 2 (June 2015): 199–224.

Gratian or two. I would extract the first- and second-recension *dicta*, those parts of the text of the *Decretum* thought to have actually been written (depending on whether one accepted Pennington's or Winroth's argument) by Gratian or by Gratian 1 and Gratian 2,[11] and run the same kind of analysis that Kestemont had run for Hildegard of Bingen and Guibert of Gembloux. I expected the results to provide an unambiguous answer, sufficiently compelling to both Pennington and Winroth to settle the debate either way, as to whether there had been one or two authors.

In August and September of 2013, I replicated the working software environment with which Kestemont had obtained his Hildegard results, installing R, R Studio, and the stylometry for R package that Kestemont had written with Maciej Eder and Jan Rybicki.[12] I started extracting text samples from Reuter and Silagi's e-text of the Friedberg edition of the *Decretum*. The fact that the e-text was encoded in the obsolete

[11] To the extent that there is some one person we can point to as corresponding to our idea of "Gratian," it's the author of the first-recension *dicta*. "The *dicta* in Gratian's *Decretum* bring the reader closer to its author than any other part of the text." Winroth, *The Making of Gratian's Decretum*, 187. **Is there anything else that can be used to support this point in "The men behind the 'Decretum'", pp.175-192?**

[12] Maciej Eder, Mike Kestemont, and Jan Rybicki, "Stylometry with R: A Suite of Tools," in *Digital Humanities 2013: Conference Abstracts* (Lincoln, NE: University of Nebraska–Lincoln, 2013), 487–89, http://dh2013.unl.edu/abstracts/ab-136.html.

(and not tree-structured) Oxford Concordance Program format made this an extremely difficult and time-consuming process. In fact, the only parts of the e-text that could both be easily extracted using Python regular expressions and, once extracted, quickly verified to be correct were the case statements. This made the case statements an obvious first choice for a test sample, although my ultimate goal was to compare only the first- and second-recension *dicta*.

Next, I needed a distraction text presumably not written by Gratian. For that purpose, I chose extracts from the pseudo-Augustinian *De vera et falsa penitentia* quoted extensively by Gratian in his *de Penitentia*, a treatise on penance inserted at C.33 q.3 in the second part of the *Decretum*. In the interest of getting fast results, I used the `vi` text editor to hand-edit the excepts directly out of the Reuter and Silagi e-text. With the case statements and the *De vera* extracts in hand, I now had enough in the way of text samples to verify that I had installed and configured R, R Studio, and stylo correctly. I have to admit that I was somewhat disappointed that the results of the first test were exactly what I should have expected: the case statements and the excerpts from *De vera* formed distinct clusters **[reproduce!]**, indicating that they were written by two different individual authors. As *De vera* is an anonymous work that predated the *Decretum* by no

more than a decade or so, and because Gratian was one of the earliest authors to quote extensively from it (although not the earliest, as I mistakenly believed at the time), I thought it would make an excellent dissertation topic if it could be shown that Gratian had forged *De vera*.

Having confirmed that my test environment could correctly distinguish the authorship of the case statements from that of the pseudo-Augustinian excerpts from *De vera*, I moved on to the much slower process of hand-editing text samples of the first- and second-recension *dicta* from the Reuter and Silagi e-text.[13]

By the second week of September 2013, I had edited the first- and second-recension *dicta* for the first part of the *Decretum* (D.1-101).

When I ran stylo on the sample, however, I got neither of the two results I had expected: either a tight clustering of all *dicta* (first- and second-recension as well as case

---

[13] For the purpose of comparing the first- and second-recension *dicta*, I define the first-recension *dicta* as the *dicta* (*ante* and *post*, but not *init.*) in the first and second parts of the Friedberg edition of the *Decretum* to which I apply the transformations defined by Winroth's appendix. I define the second-recension *dicta* as the *dicta* (*ante* and *post*, but not *init.*) in the first and second parts of Friedberg remaining after the proxy first-recension text generated by applying the Winroth transformations has been subtracted.

statements) indicating a single author and confirming all of Pennington's arguments for the unity of Gratian, or alternatively, a bimodal distribution confirming Winroth's arguments for Gratian 1 and Gratian 2. Instead, these preliminary results seemed to suggest that the first recension *dicta* had many authors, perhaps one or two of whom went on to write the second recension *dicta*. What was completely unexpected, however, was that the case statements clustered so far away from the *dicta*, extremely strong evidence that they had not been written by the same author. I immediately realized that if this accidental result held up under further testing, it would be both significant and controversial. (See Figure 1 below.)[14]

---

[14] The statistical technique of principal components analysis (PCA) projects or flattens an n-dimensional vector space representing the total variation among/between a set of samples into a more easily-visualized 2-dimensional plot. In this case, 65 vectors representing the variation in the frequency of occurrence of the 65 most frequent words in the text samples were collapsed into a smaller number of synthetic principal components. The horizontal x-axis represents the first principal component (PC1), which represents 16.9% of the total variation among/between the samples. The vertical y-axis represents the second principal component (PC2), which represents 12.5% percent of the total variation among/between the samples. The units along the x- and y-axes are standard deviations away from the means (indicated by the dashed lines) for each of the two principal components. Principal components analysis and its application to the problem of authorship attribution will be covered in depth in Chapter 4, Stylometry.

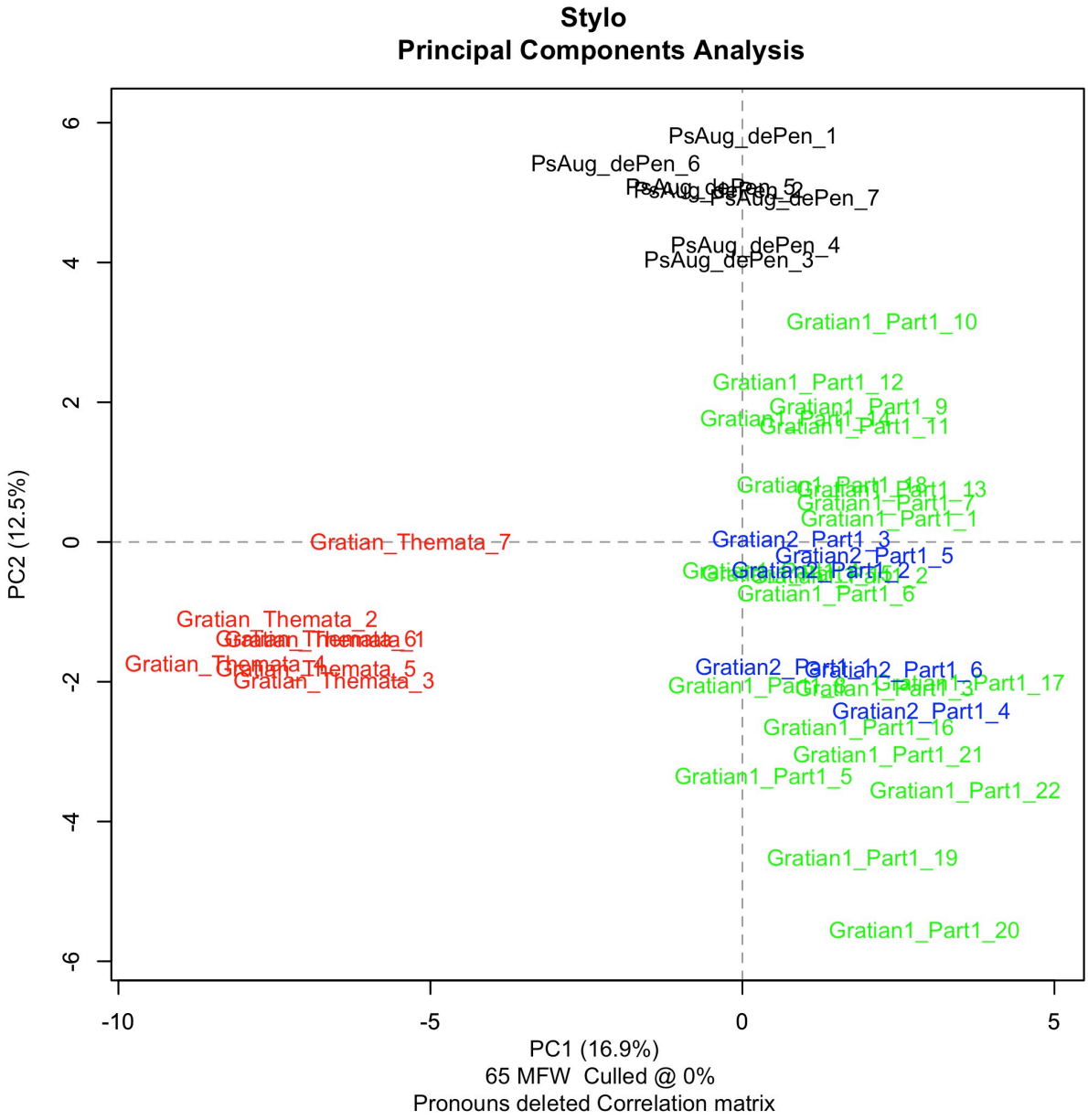

Figure 1 10 Sep 2013

Scholars working in the field have long been accustomed to thinking of the author of the *dicta* (or after Winroth's discovery, at least the author of the first-recension *dicta*) as Gratian. My initial interpretation of these surprising results was therefore that Gratian had not written the case statements. Soon, however, I came to see the JPEG image produced by stylo as telling a different and very specific "likely story"—a phrase borrowed from Plato's *Timaeus*—or what Pennington calls a "conjectural novella" about the earliest beginnings of Gratian's project, and by extension, about the dawn of the European university, the moment when the medieval school began to evolve into the faculty.

Many scholars, notably Noonan and Pennington, have seen the thirty-six cases that make up the second part of the *Decretum*, each organized around a case statement, as Gratian's unique, original, contribution to the teaching of canon law.[15] There is also a scholarly consensus foundational to most recent work on the composition of the *Decretum* that Gratian drew on just five formal sources for the bulk of the authorities he

---

[15] John T. Noonan, "Catholic Law School - A.D. 1150," *Catholic Univeristy Law Review* 47 (1997): 1201; and Kenneth Pennington, "The Biography of Gratian, the Father of Canon Law," *Villanova Law Review* 59 (2014): 689.

cited.[16] These observations prompted me to reframe my initial interpretation, and consider the possibility that the eponymous Gratian who gave his name to the entire project had written *only* the case statements.

Noonan ended his article "Gratian Slept Here" with a contemporary report of an 1143 case argued at San Marco in Venice in which a Gratian participated as a consultant to the judge. Many subsequent books and articles have referred to Noonan's discussion of the courtroom sighting of "the silent figure in the shadows of S. Marco."[17] I saw the principal components analysis plot as an indirect but compelling classroom sighting of Gratian: seated at a table with his case statements in hand and their lists of question as his syllabus, he harmonized the canons for his students directly out of the formal sources in the form of a pile of books on the table in front of him.

---

[16] Winroth, *The Making of Gratian's Decretum*, 15. Roughly one-fifth of the text of the *Decretum* has traditionally been attributed to Gratian himself; the other fourth-fifths of the text is made up of excerpts from the authorities Gratian cited.

[17] John T. Noonan, "Gratian Slept Here: The Changing Identity of the Father of the Systematic Study of Canon Law," *Traditio* 35 (January 1979): 171–72.

This conjectural novella provides a way to make sense of the fact that the author of the case statements does not appear to have written either the first- or-second recension *dicta*. In the beginning, the *Decretum* existed only in the form of the master expounding the canons to his students in a classroom presentation guided by the questions in the case statements. The overall organization, the wording of the case statements and questions, and the methodology of the *Decretum* are all Gratian's, and his students clearly thought it worthwhile to preserve the substance of his arguments, but the words are not his. The Decretum "may be a record of the first 'university course' in canon law ever taught,"[18] but the authorship attribution results suggest that we owe the written form of that record to the students rather than to their master. The strong evidence is that Gratian's direct involvement in the project came to an end, whether through death, declining health, or ecclesiastical promotion, before the first-recension *dicta* were preserved in their permanent written form.

---

[18] Winroth, *The Making of Gratian's Decretum*, 194.

Outline of Chapters

Background, the *Decretum*, Gratian, Stylometry, Next steps.

Note on the Title

University policy required me to decide on the final title of my dissertation, "Distant Reading of Gratian's *Decretum*", years before I could possibly have known what the outcome of my research was going to be. In fact, another policy actually prohibited "proceed[ing] beyond the preliminary stage in the investigation of the topic" until my dissertation proposal had been approved, but the final title still had to be submitted as part of the proposal. The "distant reading" of the title is a nod to Franco Moretti's book of the same name,[19] and refers to my early plans to use MALLET to perform unsupervised topic modeling on the first and second recensions of the *Decretum* and to identify new topics added to the second recension by comparing the results. As the project evolved and the methodological emphasis shifted from unsupervised topic modeling to stylometry using principal components analysis, the original title became

[19] Franco Moretti, *Distant Reading* (London: Verso, 2013).

obsolete. If I were to choose a title today, "Computer-aided Close Reading of Gratian's *Decretum*" would more accurately reflect the results of the project as delivered.

Note on Translations

I have, wherever possible, supplied for each Latin passage quoted the corresponding passage from a published English translation.[20] In cases where no such translation was available, or I considered the available translation seriously misleading, I have supplied my own translation, indicated with the notation (trans. PLE). Special thanks to Atria A. Larson for her suggestions regarding the translation of the *Marturi placitum*.

---

[20] Katherine Ludwig Jansen, Joanna H. Drell, and Frances Andrews, eds., *Medieval Italy: Texts in Translation*, The Middle Ages Series (Philadelphia: University of Pennsylvania Press, 2009); Robert Somerville and Bruce Clark Brasington, eds., *Prefaces to Canon Law Books in Latin Christianity: Selected Translations, 500-1245* (New Haven, Conn: Yale University Press, 1998); and Augustine Thompson and James Gordley, trans., *The Treatise on Laws: (Decretum DD. 1-20)*, Studies in Medieval and Early Modern Canon Law, v. 2 (Washington, D.C: Catholic University of America Press, 1993) have been particularly helpful resources in this regard.