



Maxim Romanov

Arabic Written Tradition & the Digital Humanities

Goethe-Universität Frankfurt am Main, March 29, 2017

*Algorithmic Analyses of Corpora
in Historical Languages*



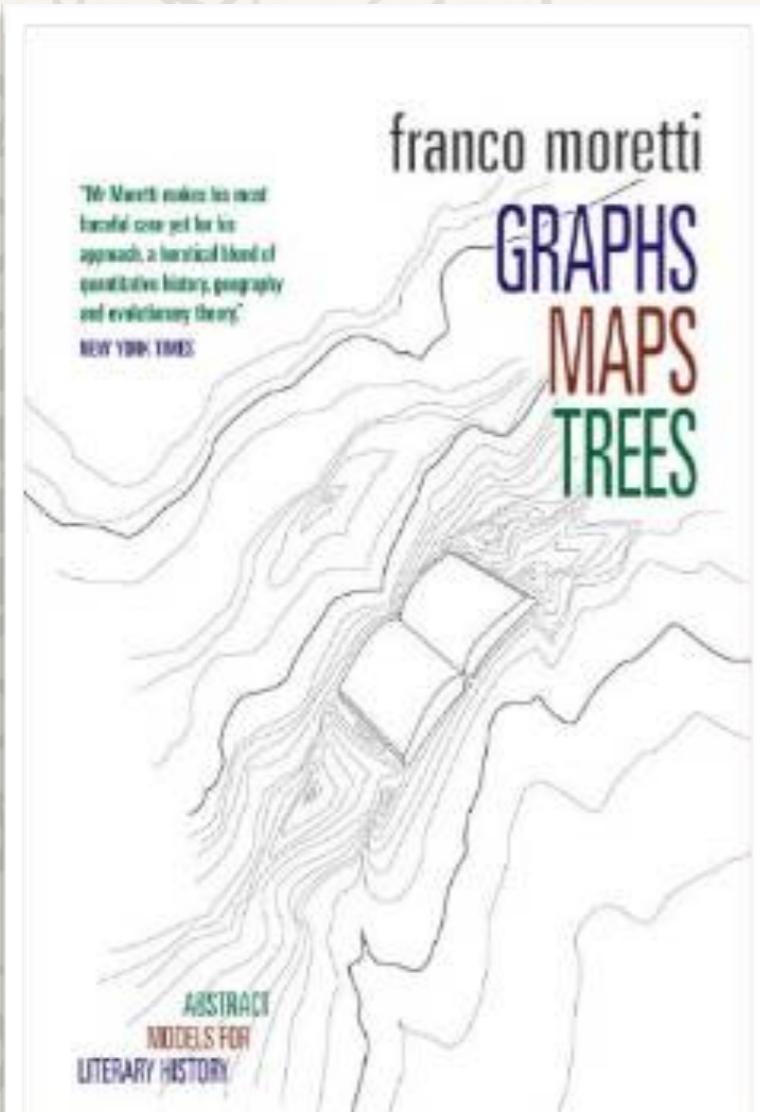
Digital Humanities
UNIVERSITÄT LEIPZIG

Digital Humanities

Engaging in a kind of humanistic research that *cannot be done without* digital methods, where complexity and novelty of each research task would vary from impossible ‘without 500 monks at hand’* to otherwise unthinkable. Such research would rely on large volumes of data (e.g., corpora of full-text primary sources) and a variety of digital approaches that allow converting raw data into meaningful information and then exploring it with different visualization techniques in order to trace long-term and large-scale developments.

* *The phrase is from:* Mathisen, Ralph W. “Where Are All the PDBs?: The Creation of Prosopographical Databases for the Ancient and Medieval Worlds.” In *Prosopography Approaches and Applications: A Handbook*, 95–126. University of Oxford, Linacre College Unit for Prosopographical Research, 2007.

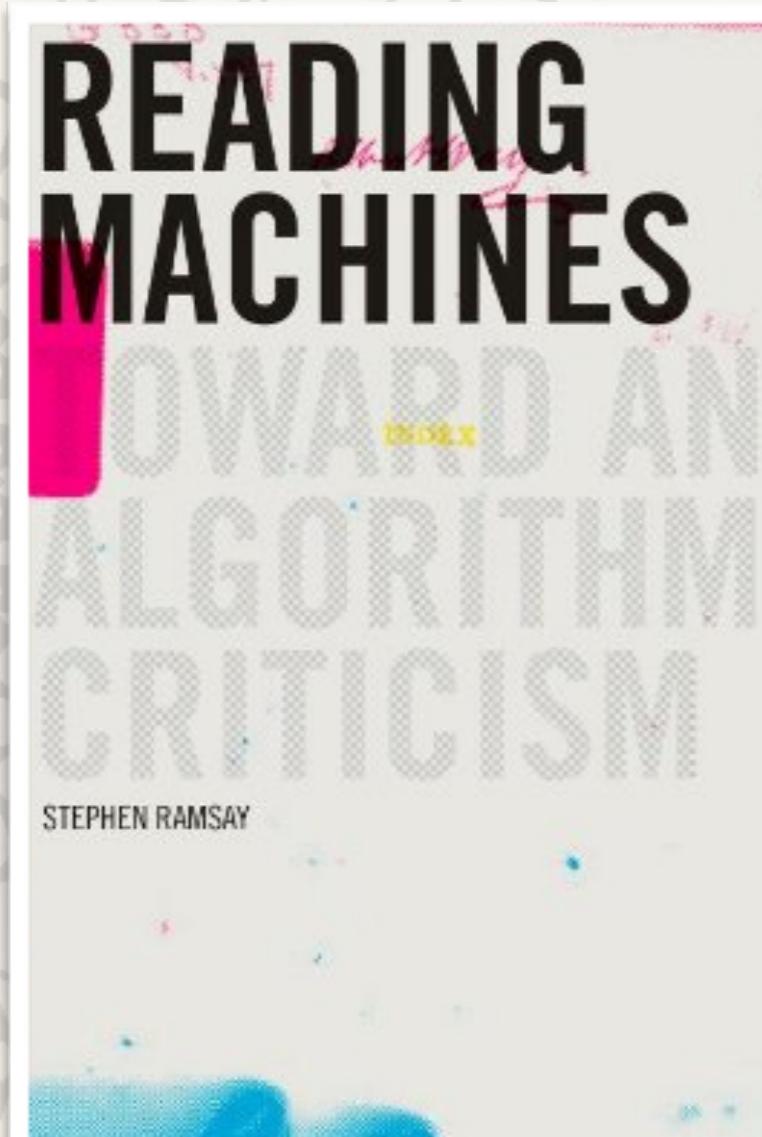
Method: *Distant Reading*



‘**Distant reading**’, I have once called this type of approach, where distance is however not an obstacle, but a specific form of knowledge: *fewer elements, hence a sharper sense of their overall interconnection*. Shapes, relations, structures. Forms. Models.

Moretti, Franco. *Graphs, Maps, Trees: Abstract Models for Literary History*. Verso, 2007.

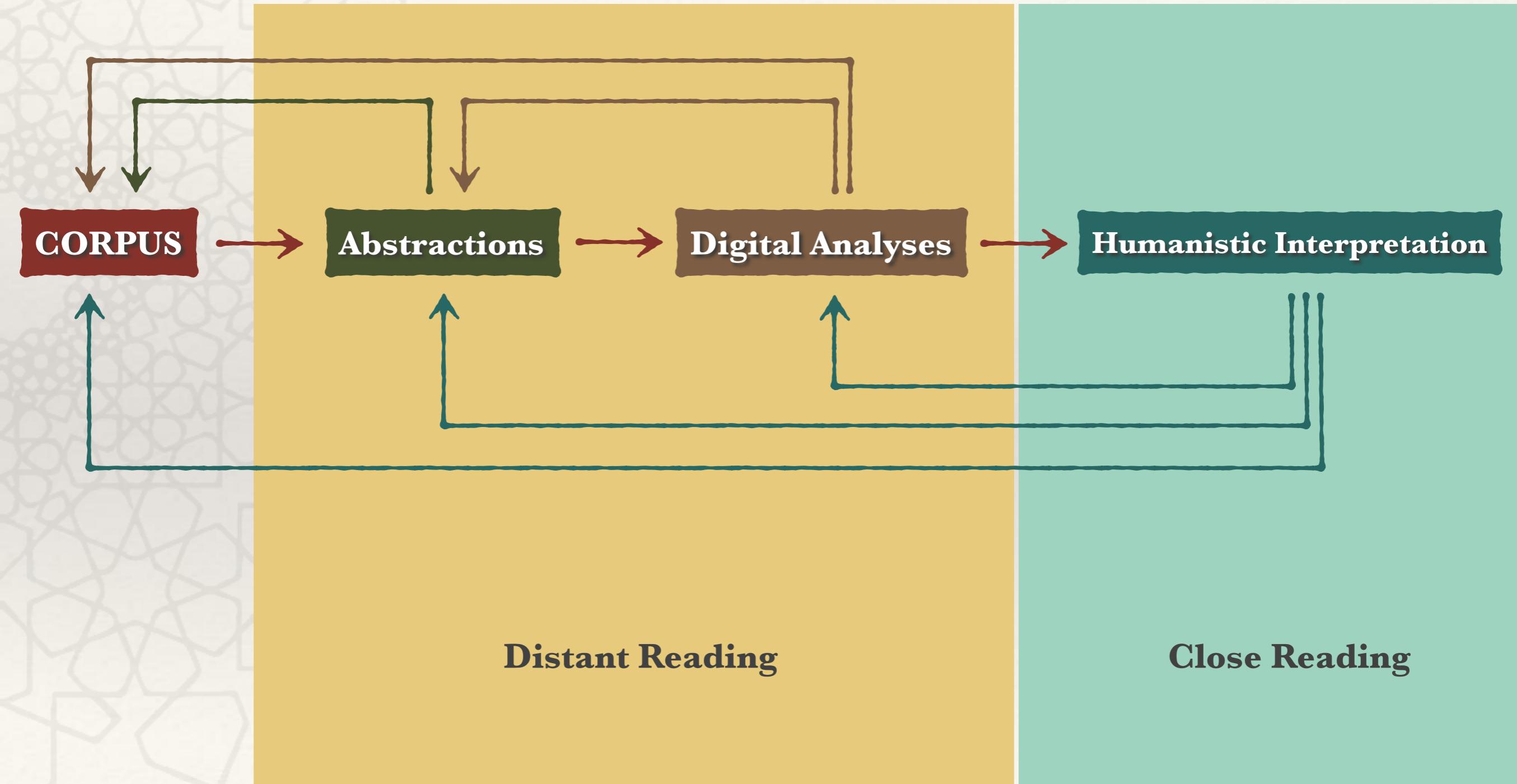
Method: *Algorithmic Deformation*



“Algorithmic criticism is easily conceived as the form of engagement that results when imperative routines are inserted into the wider constellation of texts stipulated by critical reading. But it is also to be understood as the creation of interactive programs in which readers are forced to contend not only with deformed texts, *but with the ‘how’ of those deformations.*”

Ramsay, Stephen. *Reading Machines: Toward an Algorithmic Criticism*. 1st Edition. University of Illinois Press, 2011.

Approach: *Conceptualization*



Preliminary: Corpus

Digital Collections of Arabic Texts

Library	Media	Titles	Vols	Words
<i>al-Jāmi‘ al-kabīr</i>	HDD, Windows 95	2,400	5,550	~400 mln
	www.shamela.ws	6,300	—	~820 mln
	www.almeshkat.net	7,300	—	—
	www.said.net	10,000	—	—
	www.alwaraq.com	860	—	—
<i>al-Mujam al-fiqhī</i>	DVD	1,130	3,000	—
	www.shiaonlinelibrary.com	1,970	4,175	~280 mln
<i>al-Maktaba al-shāmila</i>				
<i>al-Mishkāt</i>				
<i>Sayd al-fawā’id</i>				
<i>al-Warrāq</i>				

OpenArabic

OpenArabic

Description of the project and the status of development

[View the Project on GitHub](#)

OpenArabic/Annotation

Download
ZIP File

Download
TAR Ball

View On
GitHub

**OpenArabic Project (@ AvH
Lehrstuhl für Digital Humanities, U
Leipzig, led and curated by Maxim
Romanov)**

Contents

- [General Description](#)
- [Prospects and Progress](#)
- [Text Description Tags](#)
- [Preliminary Analysis of Categories of Texts](#)
- [Folder structure](#)
- [General description of the workflow with mARKdown](#)
- [Status Report](#)
- [List of books by centuries](#)
- [Statistics on the corpus](#)
- [Summary statistics on the lengths of texts in the corpus](#)
- [Texts by length \(duplicates excluded\)](#)
- [Texts in chronological order \(duplicates excluded\)](#)
- [Chronological Distribution of Texts - up until 1930 \(5,467 texts, 726,946,794 words\)](#)
- [Forms, Themes, Genres \(provisional assessment\)](#)

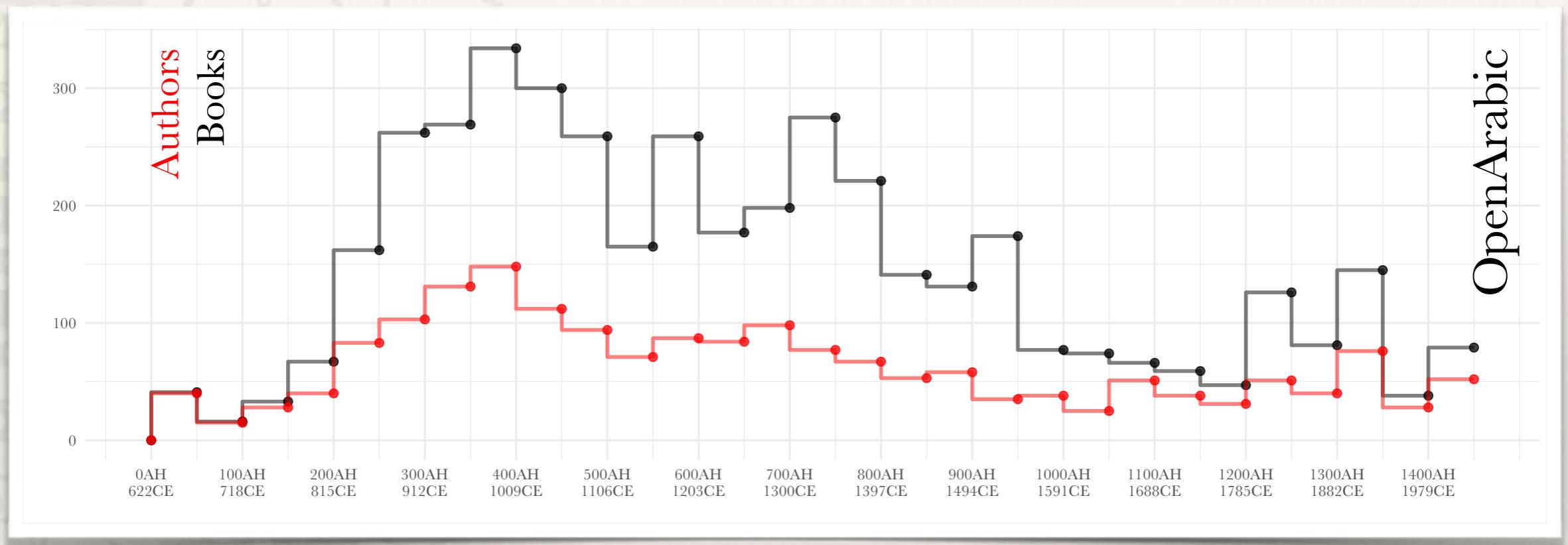
This project is maintained by [OpenArabic](#)

Hosted on GitHub Pages — Theme by [orderedlist](#)

General Description

The goal of OpenArabic is to build a machine-actionable corpus of premodern texts in Arabic to encourage computational analysis of the Arabic literary tradition. Currently, most of the texts are historical in nature (chronicles, biographical collections, geographical treatises and gazetteers,

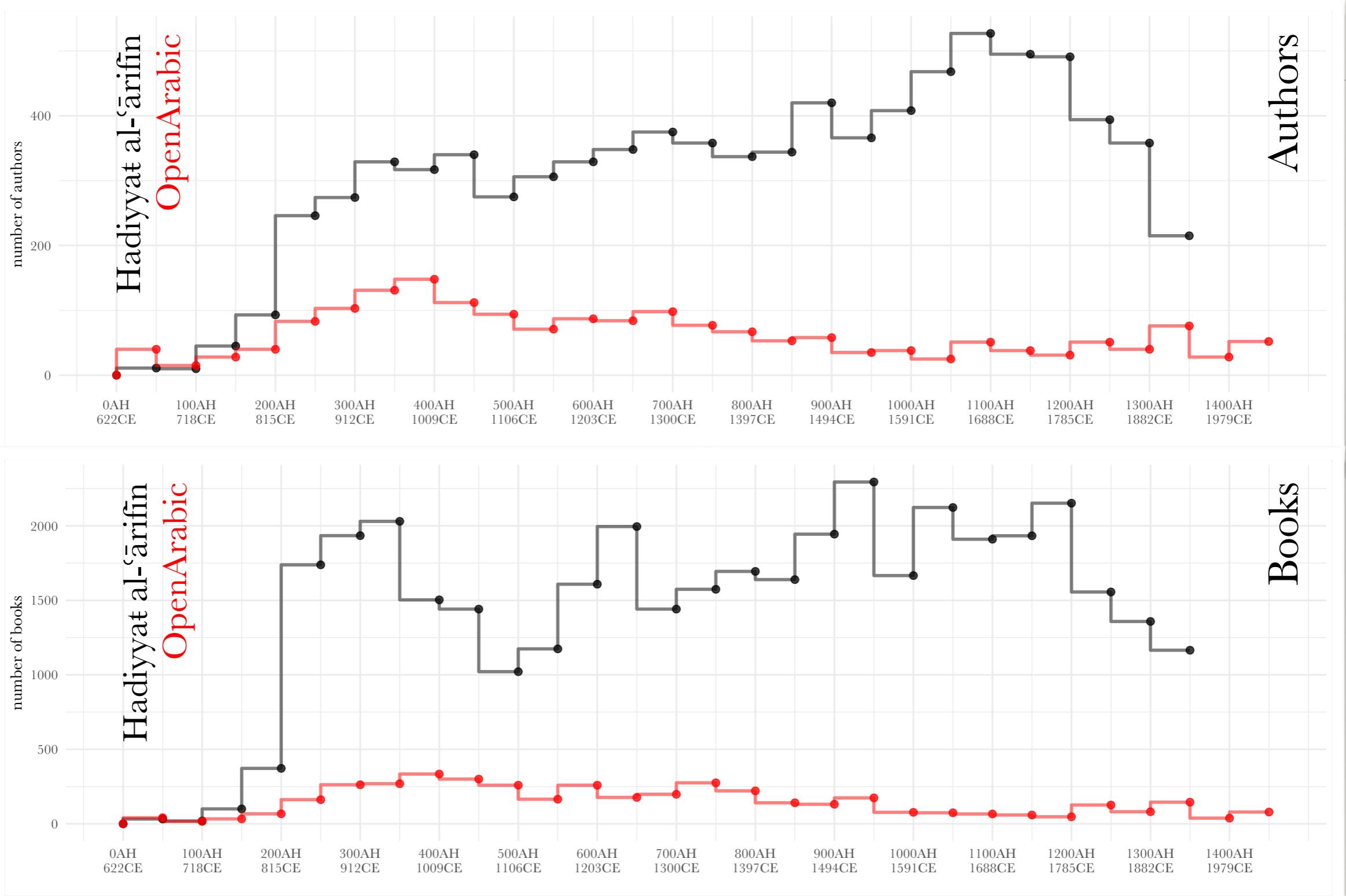
OpenArabic



**Unique Titles: 4,300
Words: 740 million
(All: 1,3 billion words)**

<https://github.com/OpenArabic/>

OpenArabic: ~10%?



Hadiyyat al-ārifīn, a bio-bibliographical collection: 8,800 authors, 40,000 titles

OpenArabic: OCR

- ❖ **Kraken ibn Ocropus (a *fork* of OCropus)**

- Benjamin Kiessling, U Leipzig
- Matthew Miller, U of Maryland
- Sarah Savant, Aga Khan U—London
- Maxim Romanov, U Leipzig

Accuracy Rates in the high 90s!



<https://www.academia.edu/28923960/>

Digital Collections: *Major Genres & Forms*

التفسير

ال الحديث

أصول الفقه ومسائله

التاريخ

الترجم والطبقات

النحو والصرف

الشعر والأدب

المعاجم

Interpretation of the Qur'ān

“Prophetic sayings”

Legal writings

Chronicles

Biographical collections

Arabic language, grammar & morphology

Poetry & fine literature

Various dictionaries & references

Digital Collections: *Major Genres & Forms*

التفسير

Interpretation of the Qur'ān

الحديث

“Prophetic sayings”

أصول الفقه ومسائله

Legal writings

التاريخ

Chronicles

الترجم والطبقات

Biographical collections

النحو والصرف

Arabic language, grammar & morphology

الشعر والأدب

Poetry & fine literature

المعاجم

Various dictionaries & references

Digital Collections: *Biographical Collections*

~300-400?

~250

structure
coverage

biographies

> 400,000

have been written

available in digital format

chronological, generational, alphabetical
social and religious, geographical

well-structured (linguistic formulae):
onomastic section, birth [date &] place,
teachers, students, contribution,
miscellanea, dear: date & place, reaction of
the community

biographies?

Case Study I: *Modeling Social History*

al-Dahabī (d. 1347 CE)
and his *Ta'rīh al-islām*
("The History of Islam")

al-Dahabī (d. 748/1347 CE)

- ❖ a Damascene religious scholar, Ḡāfi‘ī jurist and historian
 - ❖ “The History of Islam” (*Ta’rīħ al-islām*)
 - ❖ “first draft” in 714/1314 CE
 - ❖ 50 volumes (here: 4-50)
 - ❖ 7 centuries (here: 41-700/661-1300 CE)
 - ❖ over 30,000 biographies (here: ~29,100)
 - ❖ ~ 3,2 mln. words

“The History of Islam”: Distant Reading

- ❖ Death dates
- ❖ Place names / Toponyms
- ❖ “**Descriptive names**” (*nisbats*)
 - ❖ 700 unique (frequency 10 and higher)
 - ❖ 70,000 total
 - ❖ ~ social profile (issue with their meaning*)

“The History of Islam”: Distant Reading

Traditional Arab Name: Example

‘Abd al-Rahmān ibn ‘Alī ibn Muḥammad ibn ‘Ubayd Allāh ibn ‘Abd Allāh ibn Ḥamadá ibn Muḥammad ibn Ğa‘far ibn ‘Abd Allāh ibn al-Ḳāsim ibn al-Naḍar ibn al-Ḳāsim ibn Muḥammad ibn ‘Abd Allāh ibn ‘Abd al-Rahmān ibn al-Ḳāsim ibn Muḥammad ibn Abī Bakr al-Ṣiddīk (*may God be pleased with him!*) al-Ḳurašī al-Taymī al-Bakrī al-Bağdādī al-Ḥāfiẓ al-Mufassir al-Ḥanbalī al-Wā‘iz al-Ṣaffār, Ğamāl al-Dīn, Abū-l-Farağ, known as Ibn al-Ğawzī (d. 597/1201 ce)

“The History of Islam”: Distant Reading

Traditional Arab Name: Example

‘Abd al-Rahmān ibn ‘Alī ibn Muḥammad ibn ‘Ubayd Allāh ibn ‘Abd Allāh ibn Ḥamadá ibn Muḥammad ibn Ğa‘far ibn ‘Abd Allāh ibn al-Ķāsim ibn al-Naḍar ibn al-Ķāsim ibn Muḥammad ibn ‘Abd Allāh ibn ‘Abd al-Rahmān ibn al-Ķāsim ibn Muḥammad ibn Abī Bakr al-Šiddīk (*may God be pleased with him!*) al-Ķurašī al-Taymī al-Bakrī al-Bağdādī al-Ḩafiz̄ al-Mufassir al-Ḩanbalī al-Wā‘iz̄ al-Şaffār, Ğamāl al-Dīn, Abū-l-Farağ̄, known as Ibn al-Ğawzī (d. 597/1201 ce)

“The History of Islam”: Distant Reading

Traditional Arab Name: Example

‘Abd al-Rahmān ibn ‘Alī ibn Muḥammad ibn ‘Ubayd Allāh ibn ‘Abd Allāh ibn Ḥamadá ibn Muḥammad ibn Ğa‘far ibn ‘Abd Allāh ibn al-Ḳāsim ibn al-Naḍar ibn al-Ḳāsim ibn Muḥammad ibn ‘Abd Allāh ibn ‘Abd al-Rahmān ibn al-Ḳāsim ibn Muḥammad ibn Abī Bakr al-Ṣiddīk (*may God be pleased with him!*) al-Ḳurašī al-Taymī al-Bakrī al-Bağdādī al-Ḥāfiẓ al-Mufassir al-Ḥanbalī al-Wā‘iz al-Ṣaffār, Ğamāl al-Dīn, Abū-l-Farağ, known as Ibn al-Ğawzī (d. 597/1201 ce)

“The History of Islam”: Distant Reading

Traditional Arab Name: Example

‘Abd al-Rahmān ibn ‘Alī ibn Muḥammad ibn ‘Ubayd Allāh ibn ‘Abd Allāh ibn Ḥamadá ibn Muḥammad ibn Ğa‘far ibn ‘Abd Allāh ibn al-Ķāsim ibn al-Naḍar ibn al-Ķāsim ibn Muḥammad ibn ‘Abd Allāh ibn ‘Abd al-Rahmān ibn al-Ķāsim ibn Muḥammad ibn Abī Bakr al-Šiddīk (*may God be pleased with him!*) al-Ķurašī al-Taymī al-Bakrī al-Bağdādī al-Ḩafiz̄ al-Mufassir al-Ḩanbaļī al-Wā‘iz̄ al-Şaffār, Ğamāl al-Dīn, Abū-l-Farağ, known as Ibn al-Ğawzī (d. 597/1201 ce)

“The History of Islam”: Distant Reading

Traditional Arab Name: Example

LAQAB

KUNNYAT

ŞUHRAT

‘Abd al-Rahmān ibn ‘Alī ibn Muḥammad ibn
‘Ubayd Allāh ibn ‘Abd Allāh ibn Ḥamadá ibn
Muḥammad ibn Ğa‘far ibn ‘Abd Allāh ibn al-
Kāsim ibn al-Naḍar ibn al-Kāsim ibn Muḥammad
ibn ‘Abd Allāh ibn ‘Abd al-Rahmān ibn al-Kāsim
ibn Muḥammad ibn Abī Bakr al-Šiddīk (*may God be
pleased with him!*) al-Kurašī al-Taymī al-Bakrī al-
Bağdādī al-Ḥafız al-Mufassir al-Ḥanbalī al-Wā‘iz
al-Saffār, Ğamāl al-Dīn, Abū-l-Farağ, known as Ibn
al-Ğawzī (d. 597/1201 ce)

“The History of Islam”: Distant Reading

Traditional Arab Name: Example

... *al-Kurašī al-Taymī al-Bakrī al-Bagdādī al-Hāfiẓ al-Mufassir al-Hanbalī al-Wā'iẓ al-Saffār* ...

NISBATS

al-Kurašī

member of the tribe of Quraysh (tribal)

al-Taymī

member of the clan of Taym (tribal)

al-Bakrī

descendant of Abū Bakr al-Šiddīq (ancestral)

al-Bagdādī

resident/native of Baghdād (toponymic)

al-Hāfiẓ

“Preserver” of the Tradition (religious specialization)

al-Mufassir

exegete of the Qur’ān (religious specialization)

al-Hanbalī

jurist of the Ḥanbalī legal school (religious affiliation)

al-Wā'iẓ

public preacher (religious specialization)

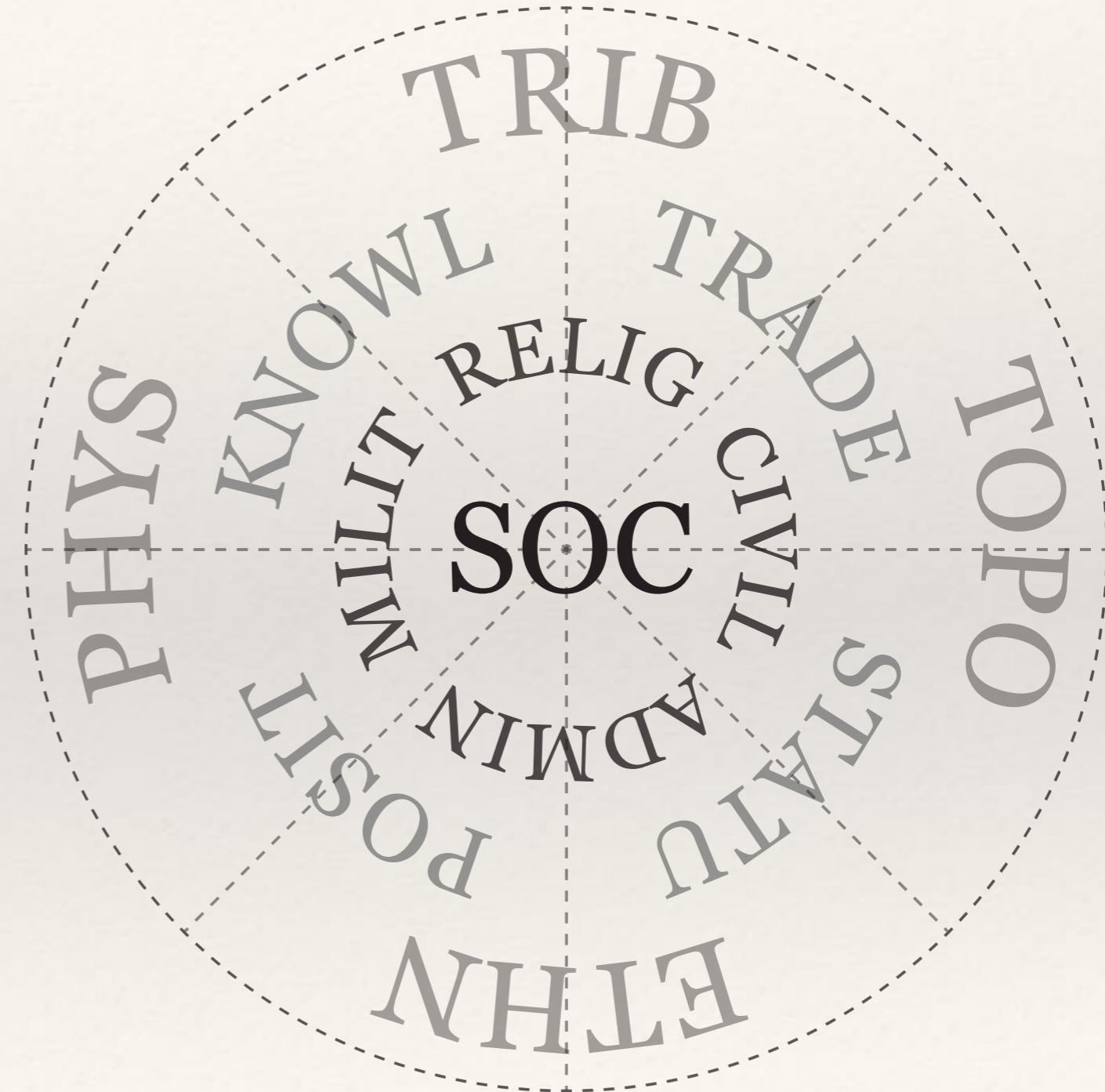
al-Saffār

seller of copper/brass utensils (occupational)

Classification of *Nisbats* from the Social Perspective

Previous work on *nisbats*
(mostly occupational):

- Hayyim Cohen
- Carl Petry
- Maya Shatzmiller



Algorithmic Analysis

Biographical Collections: *Structures*

L C F R	\$ # الهروي	237
L C F R	# أبو سعيد إبراهيم بن طهمان بن شعيب من قرية باشان نزيل سنابور	238
L C F R	سافر إلى مكة ومات بها كان فقيها محدثاً توفى سنة 163	239
L C F R	ومنها صنف تفسير القرآن وكتاب الفقه وكتاب العبددين	240
L C F R	المناقب	241

Biographical Collections: *Structures*

L C F R	\$. الهروي	# 237
L C F R	أبو سعيد إبراهيم بن طهمان بن شعيب من قرية باشان نزيل نيسابور	# 238
L C F R	سافر إلى مكة ومات بها كان فقيها محدثاً توفى سنة 163 ثلاث وستين	~~ 239
L C F R	ومائة . صنف تفسير القرآن . سنن الفقه . كتاب العبددين . كتاب المناقب	~~ 240
L C F R	المناقب .	~~ 241

\$ al-Harawī

Abū Sa‘īd Ibrāhīm b. Ṭahmān b. Šu‘ayb, from the village of Bashan, a resident of Nishapur.

~~ He traveled to Mecca and died there. He was a jurist, transmitter of Hadith. He died in 163.

~~ ... **He composed** The Exegesis of the Qur'an, Legal hadith, The Book of Two Celebrations, The Book of Virtues.

Biographical Collections: *Structures*

17 ॥### \$ - Harawī .
18 # Abū · Sa'īd · Ibrāhīm · ibn · Ṭahmān · ibn · Šu'ayb · @S01 · Harawī , · from · the · village ·
19 ~~of · @T01 · Bāšān , · a · resident · of · @T01 · Naysabūr · [Nishapur] . · He · traveled · to ·
20 ~~@T01 · Makkat · [Mecca] · and · died · there . · He · was · a · @S01 · jurist · and ·
21 ~~a · @S01 · traditionist . · He · died · in · the · @YD163 · year · one · hundred · sixty · three · .
22 ~~He · wrote · : · Tafsīr · al-Qur'ān , · Sunan · al-fiqh ,
23 ~~Kitāb · al-'īdayn , · Kitāb · al-manāqib .



=====

id, item, category

=====

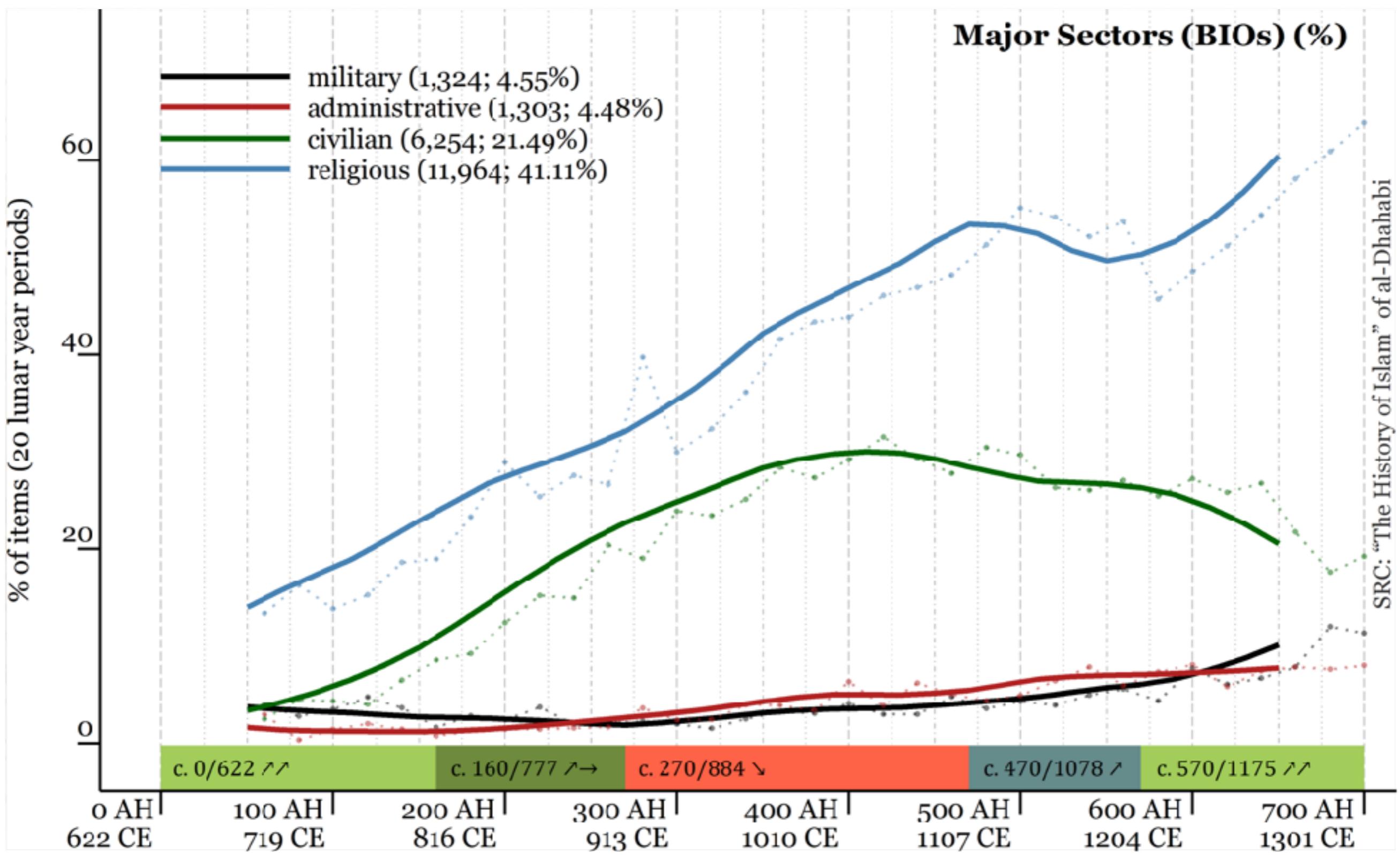
000006, 163, year_of_death
000006, Bāšān, toponym
000006, Naysabūr, toponym
000006, Makkat, toponym
000006, Harawī, descriptive_name
000006, jurist, descriptive_name
000006, traditionist, descriptive_name

=====

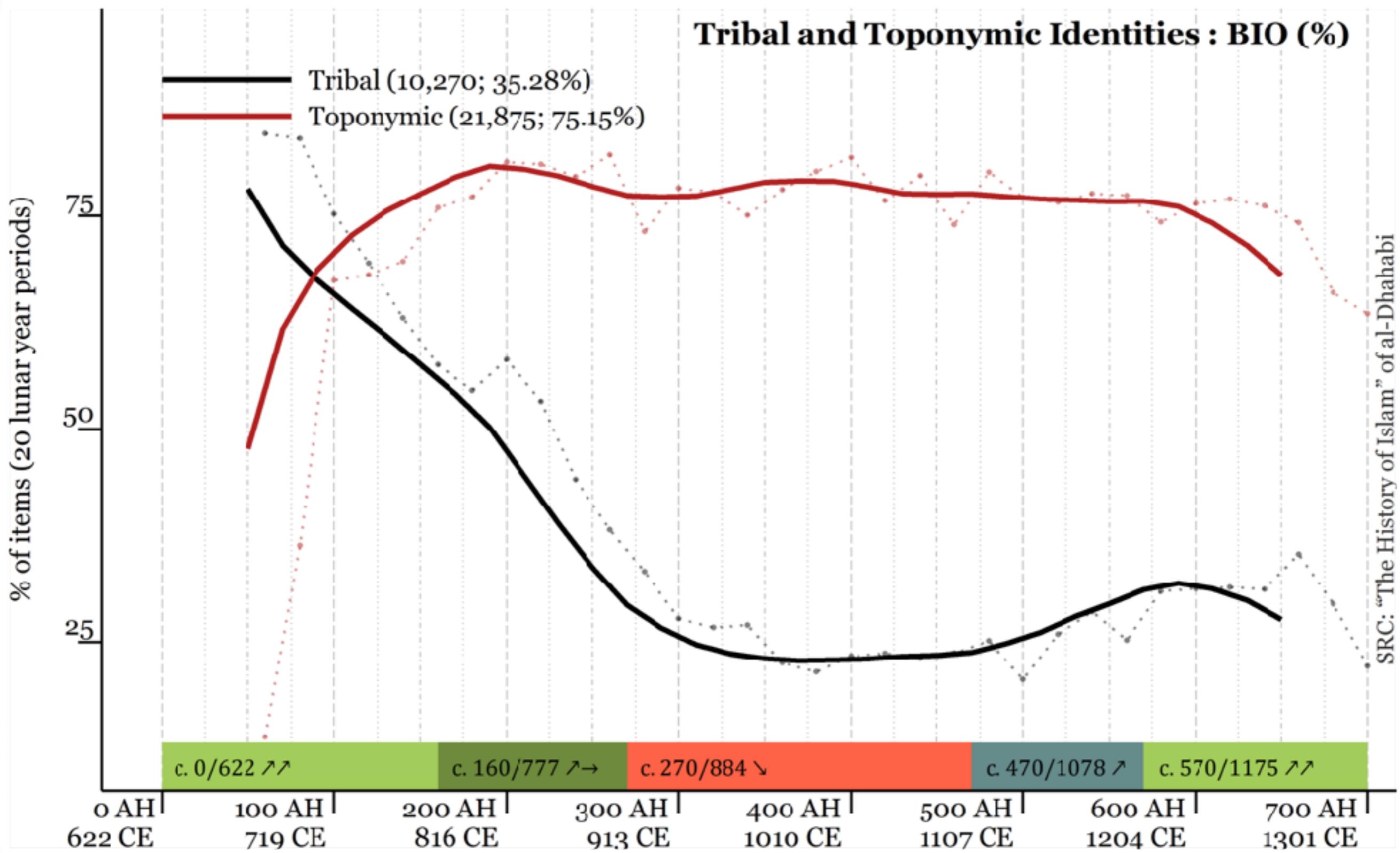


Social transformations

Major Sectors (BIOS) (%)



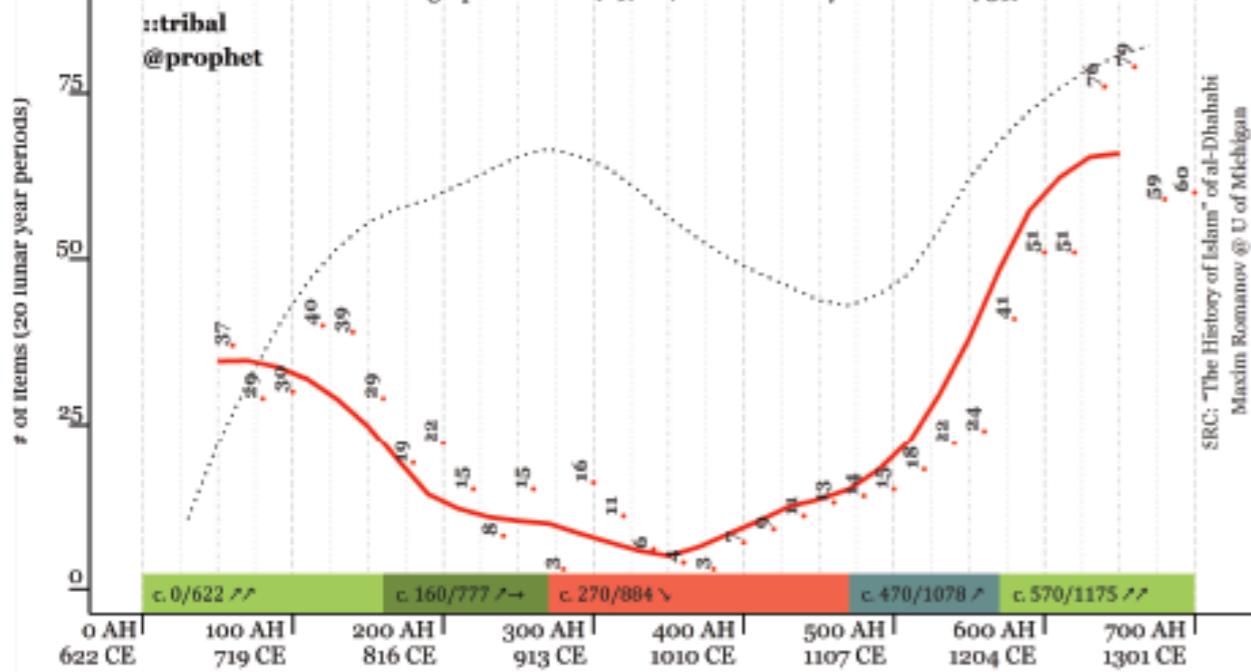
De-tribalization



De-tribalization

ANSARI (876 total; BW: AnSAry)

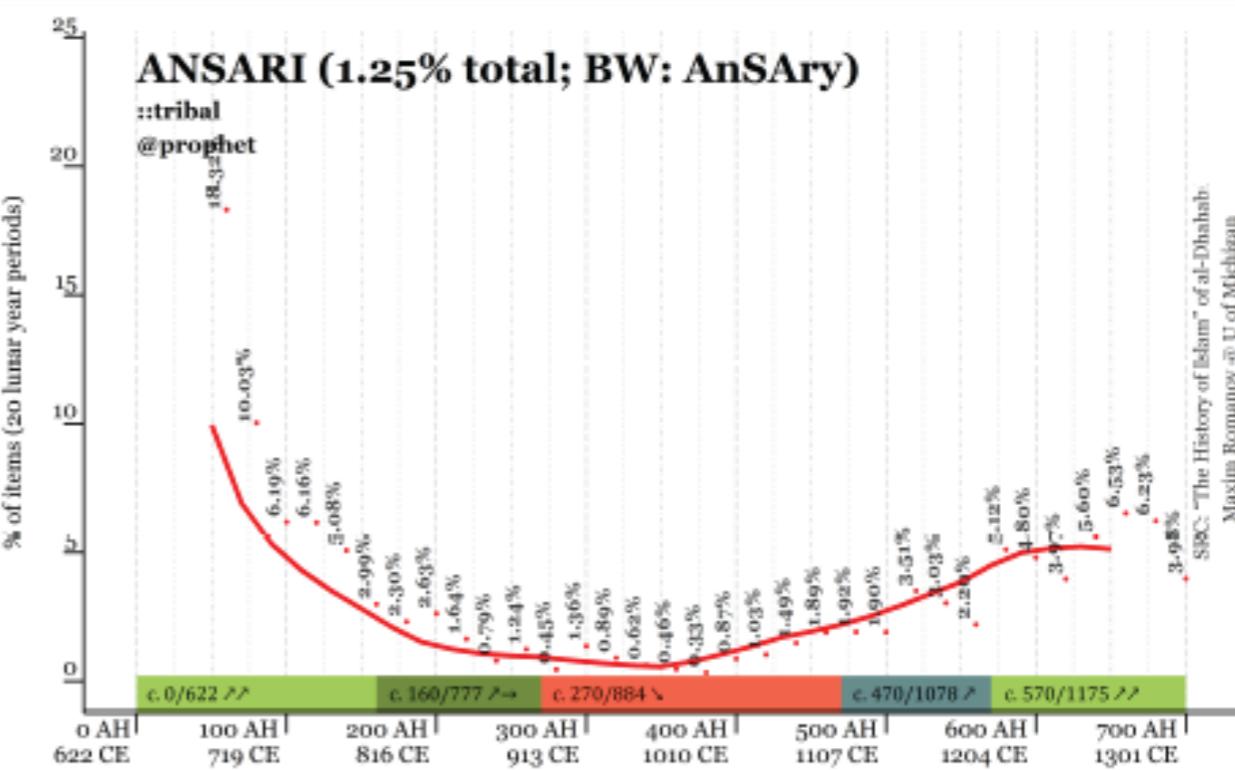
Cumulative biographical curve (29,110; scaled down by a factor of 14.59)



ANSARI (1.25% total; BW: AnSAry)

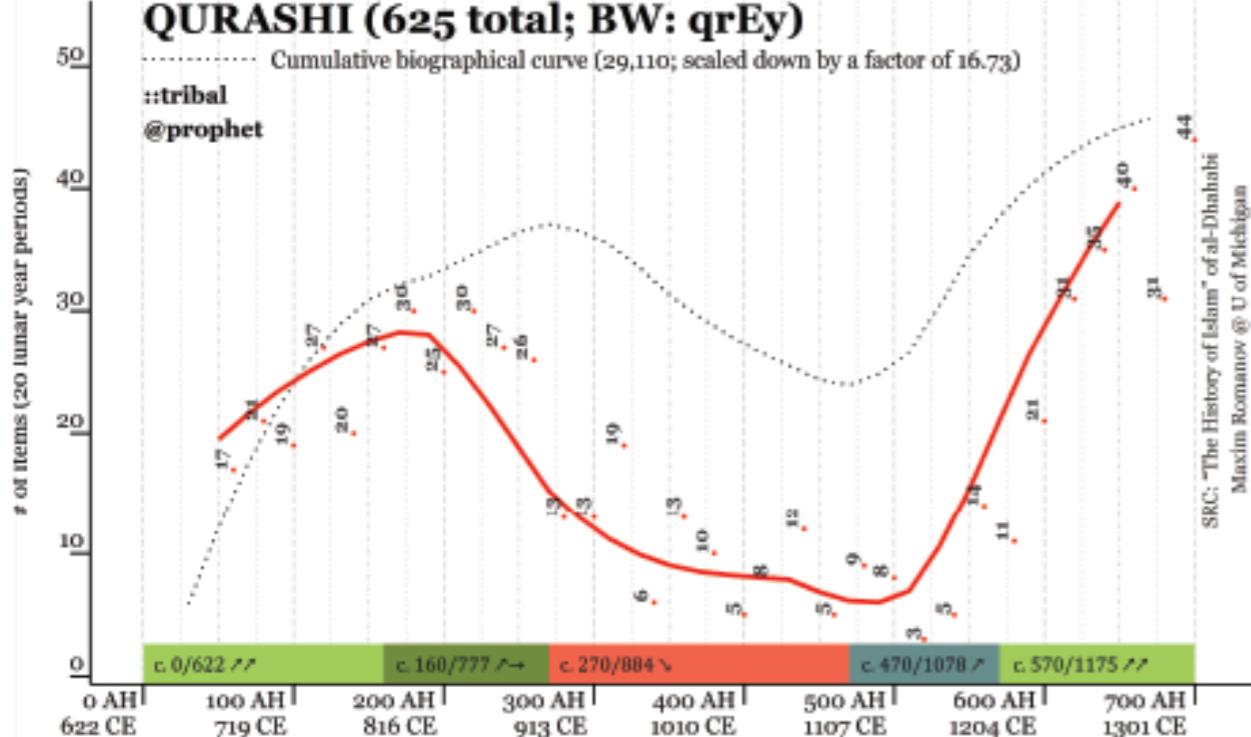
::tribal
@prophet

SRC: "The History of Islam" of al-Dhahabi
Maxim Romanov © U of Michigan



QURASHI (625 total; BW: qrEy)

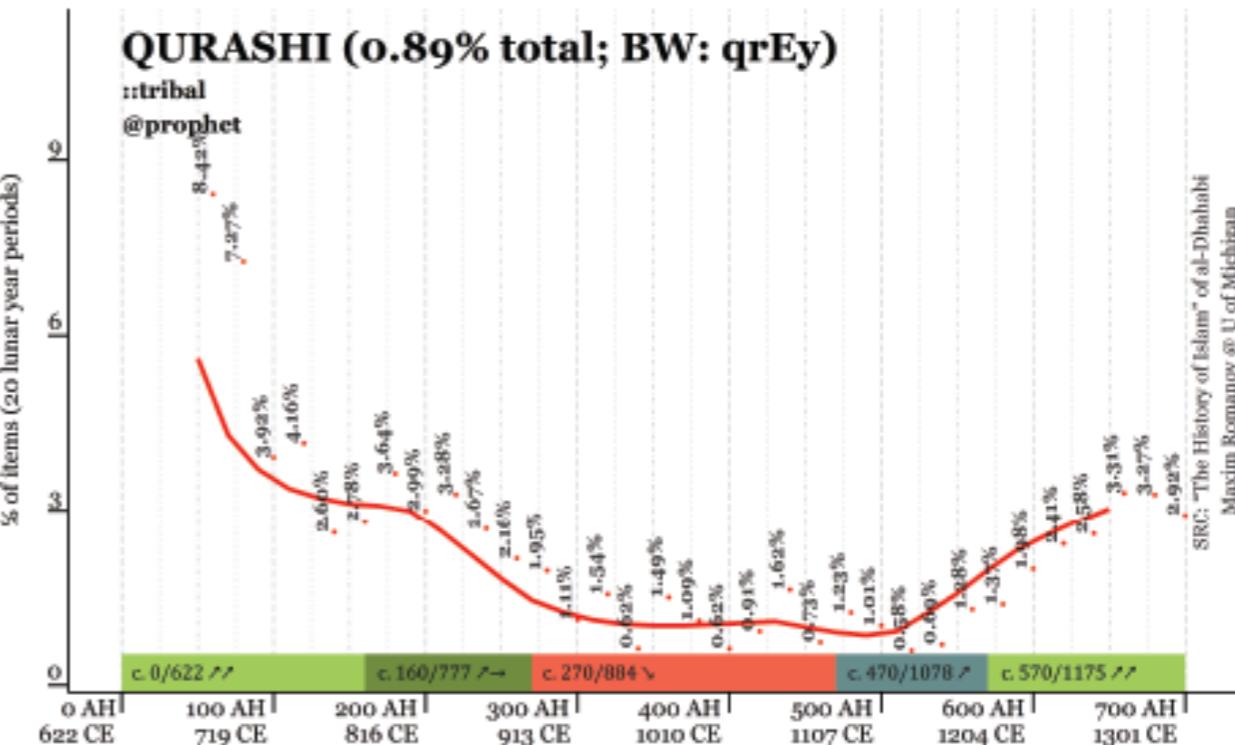
Cumulative biographical curve (29,110; scaled down by a factor of 16.73)



QURASHI (0.89% total; BW: qrEy)

::tribal
@prophet

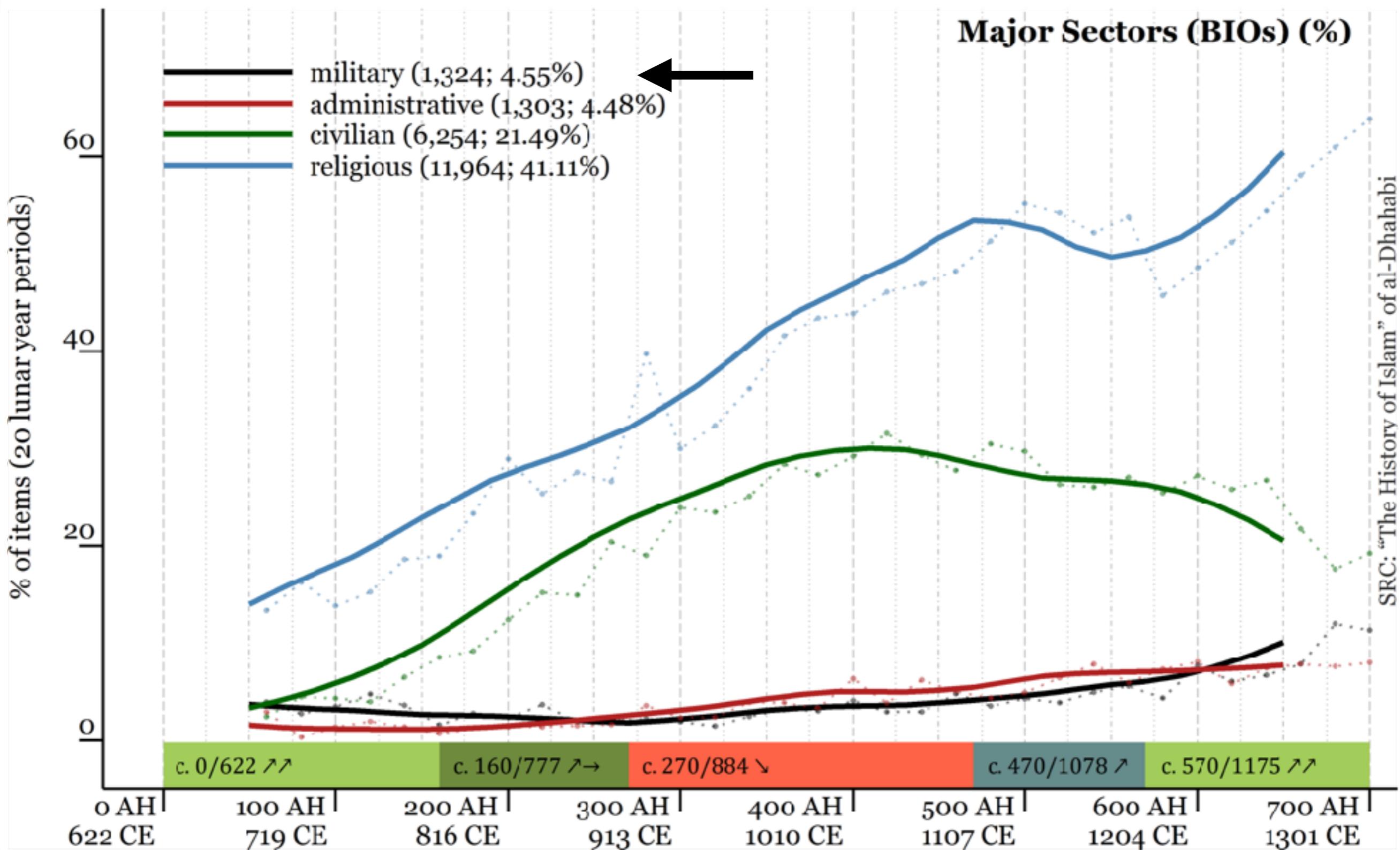
SRC: "The History of Islam" of al-Dhahabi
Maxim Romanov © U of Michigan



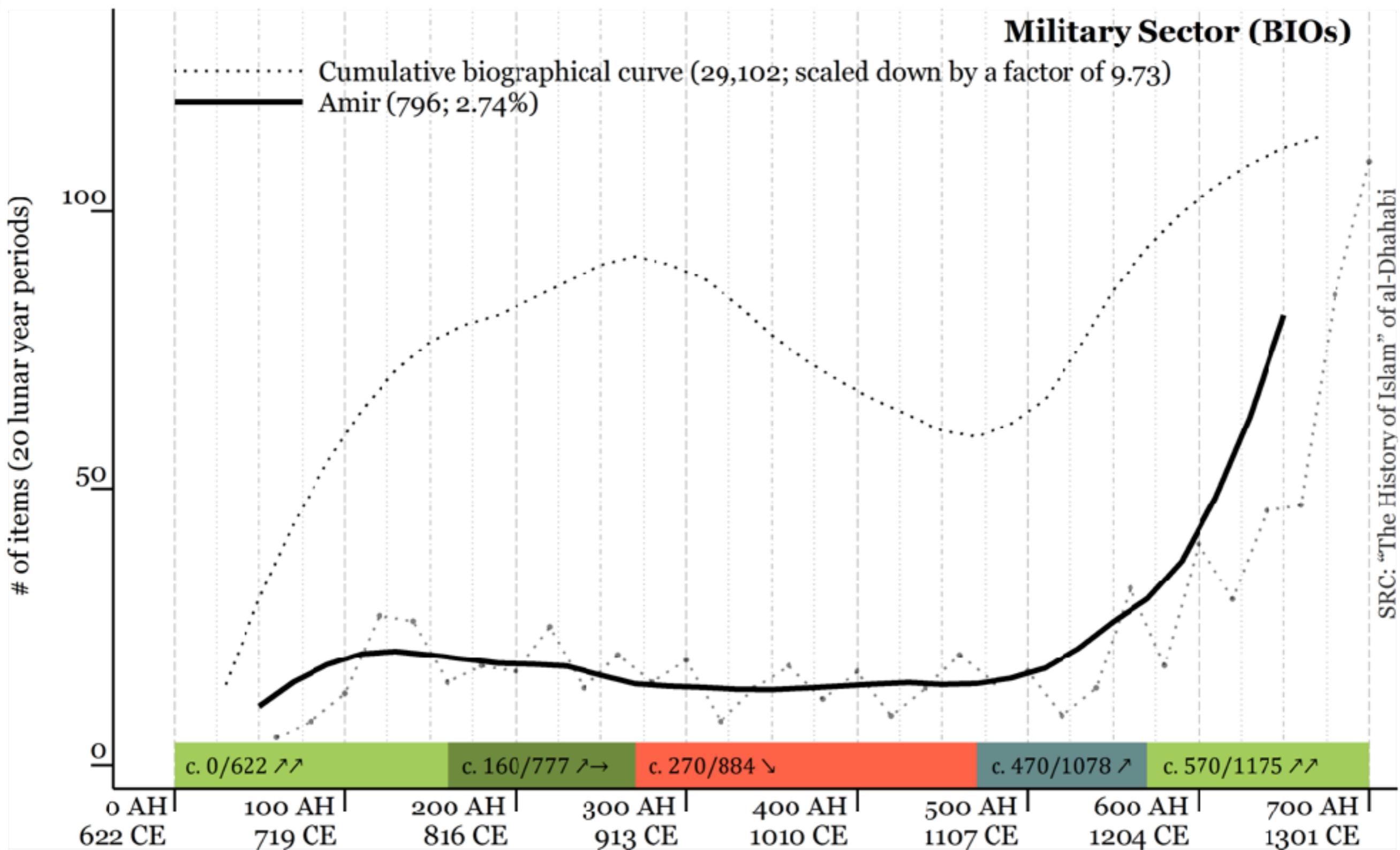
“The reduction of central political authority took the form of what was perhaps the most distinctive feature of the Middle Islamic periods in general, as compared with other times and areas in the Agrarian Age:
its militarization.”

—*Marshal Hodgson, The Venture of Islam. Vol. II, p.64*

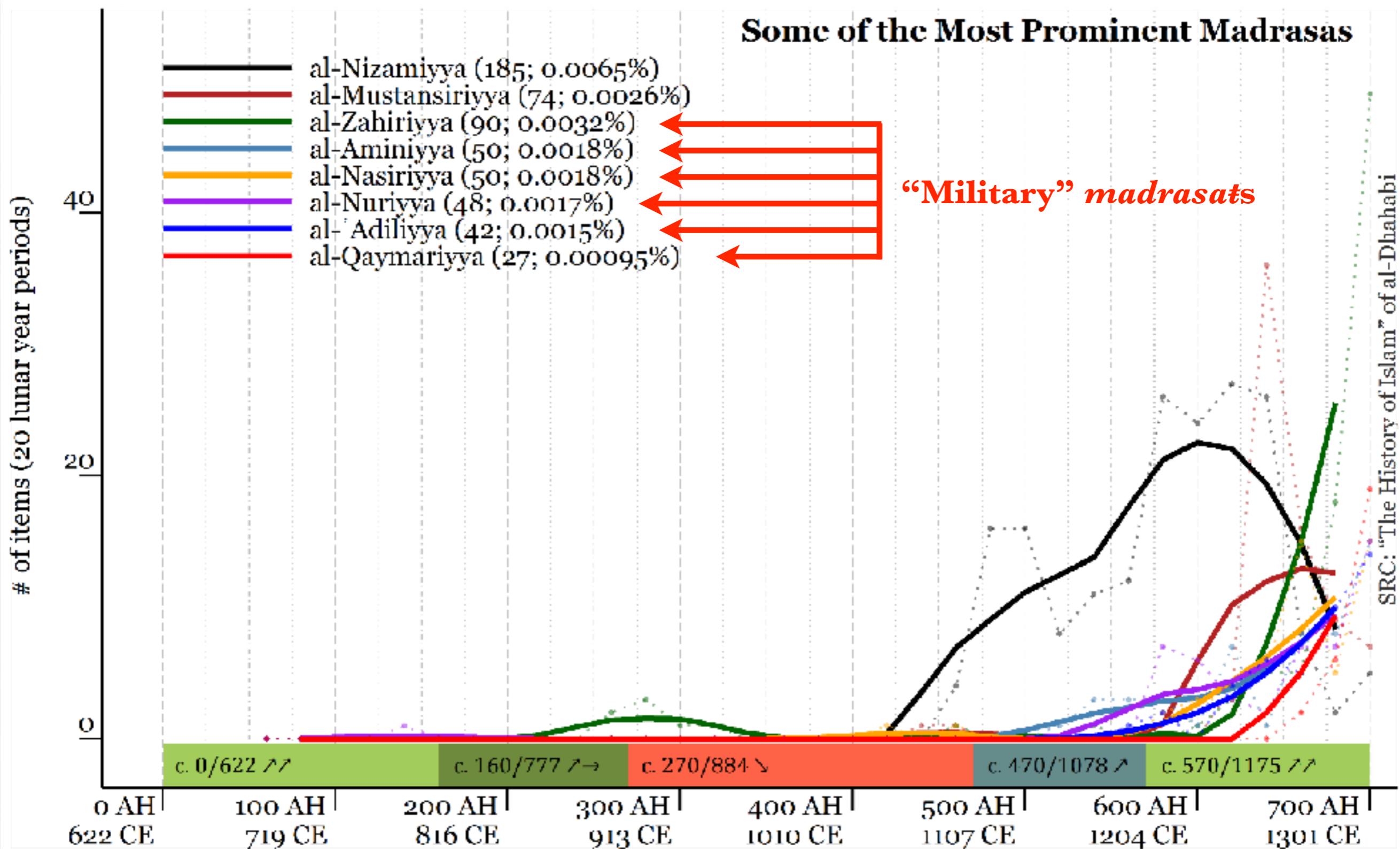
Militarization of élites



Militarization of élites

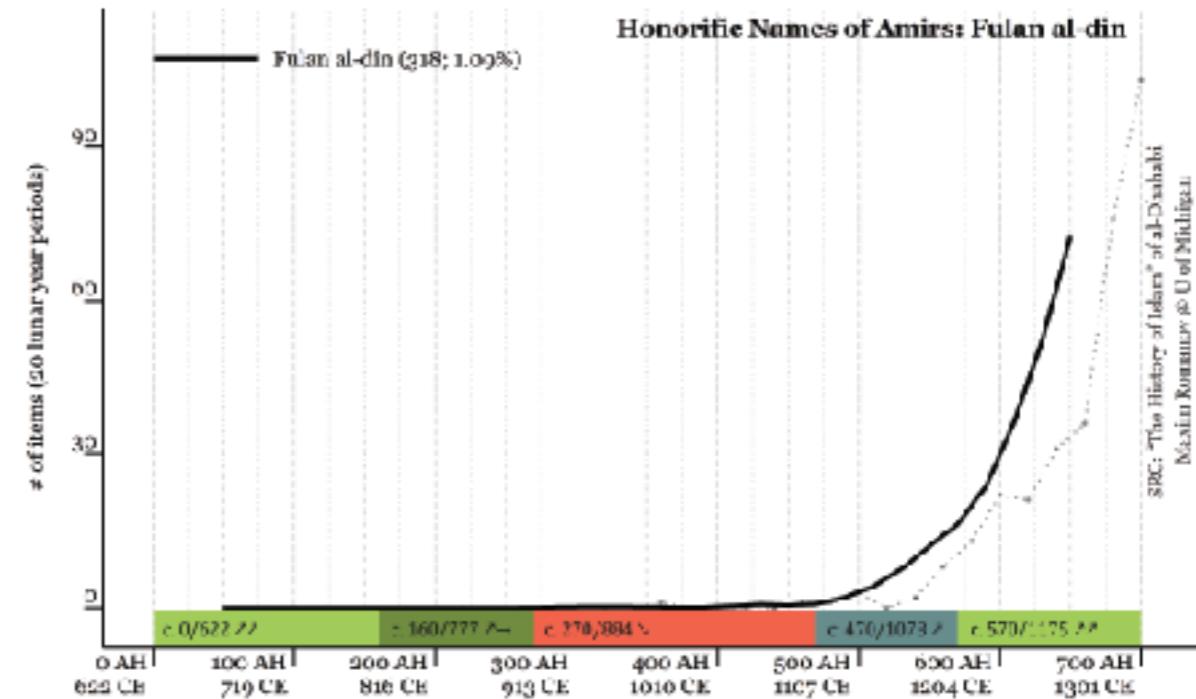
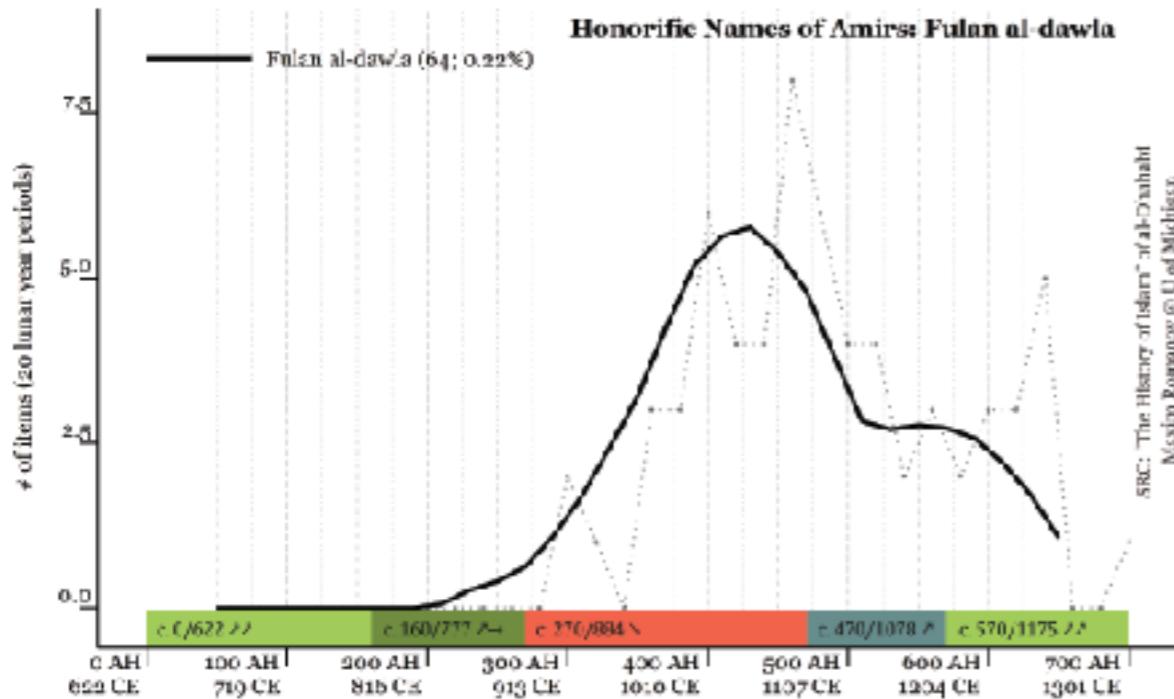


Militarization of élites



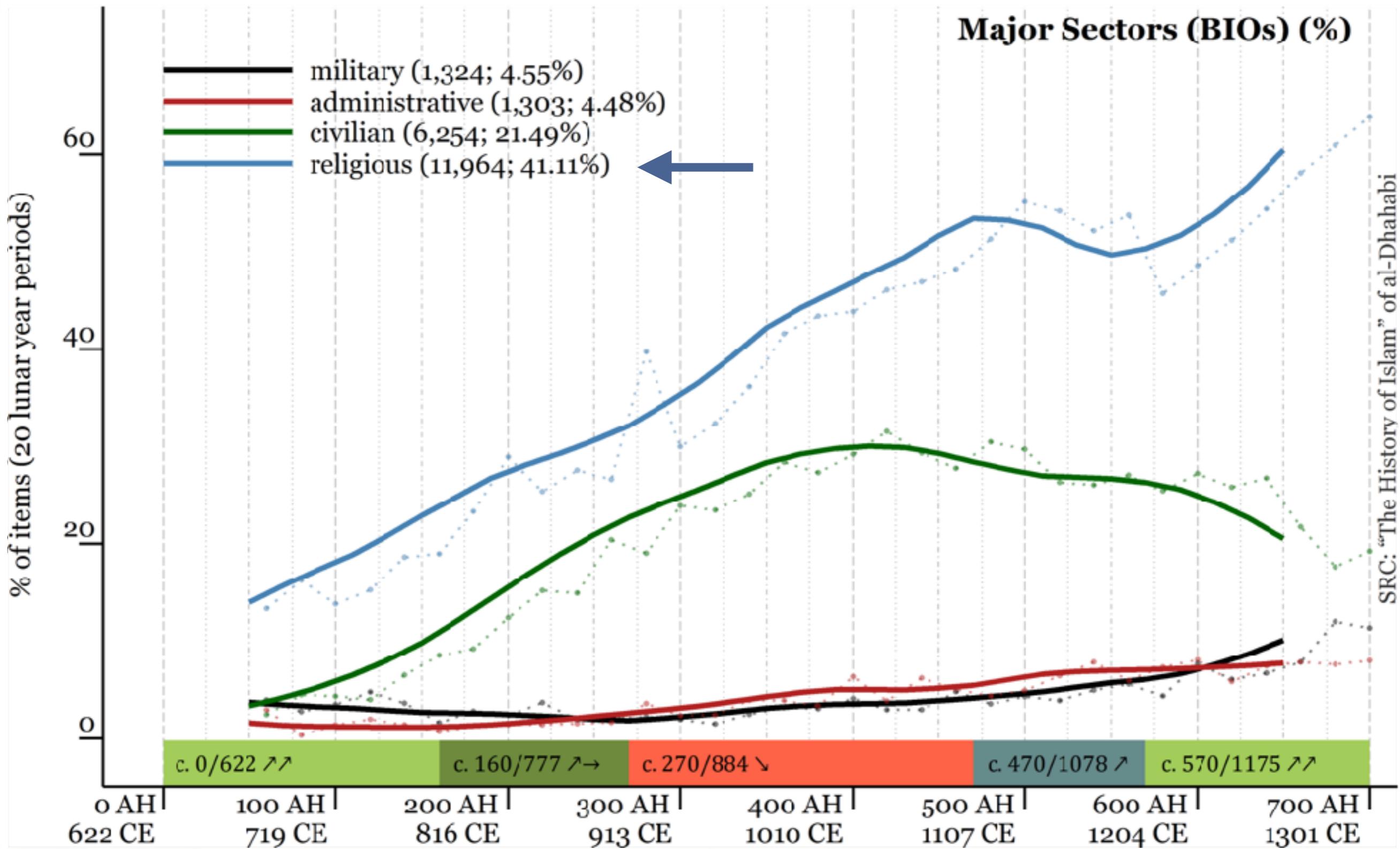
Militarization of élites: Honorific Names

Sayf al-dawlat > Sayf al-dīn “Sword of the Dynasty” > “Sword of Religion”

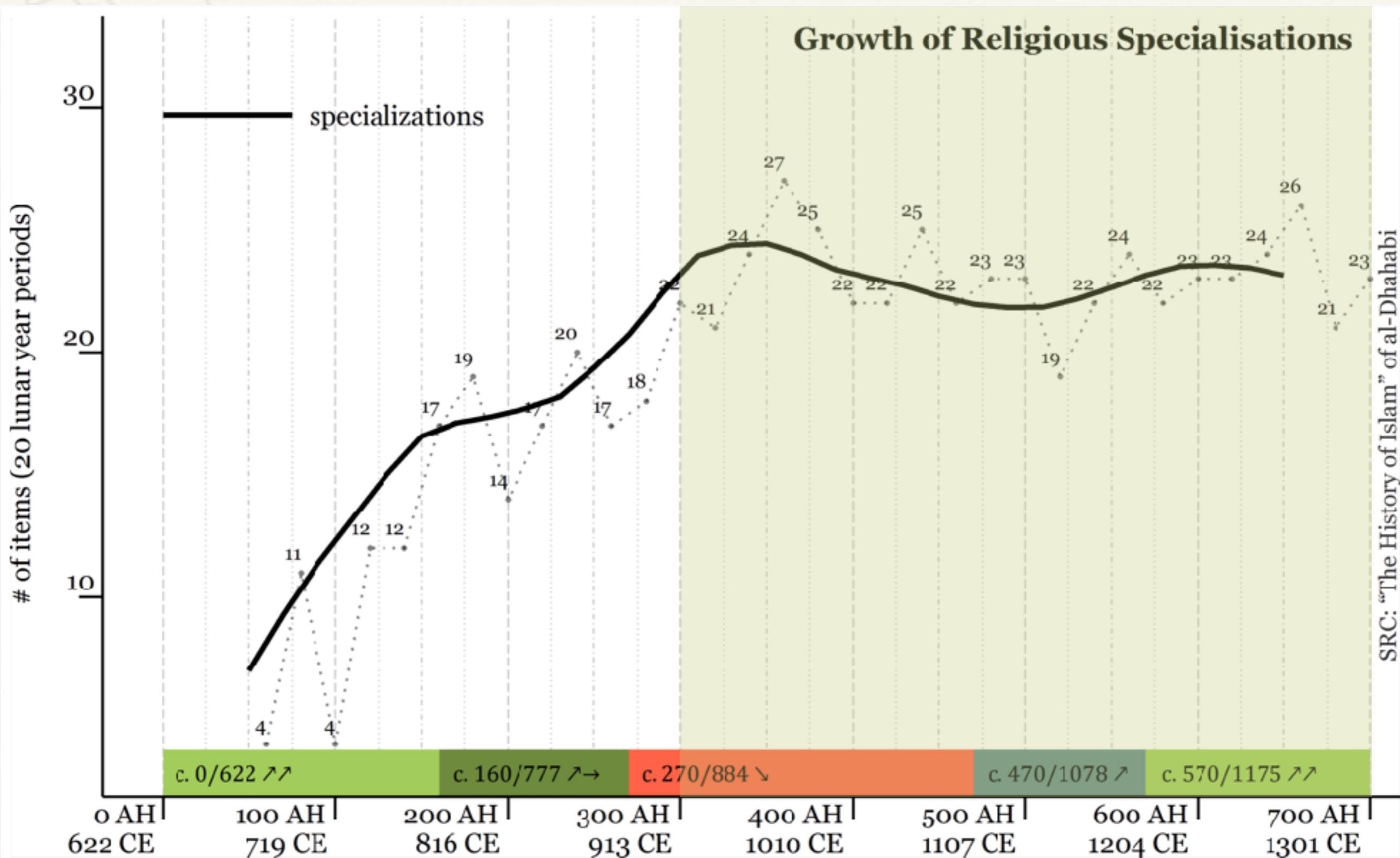


Patterns of Military Honorific Names: Fulān al-dawla, the most common pattern in the middle period, gets replaced by Fulān al-dīn pattern in the later period

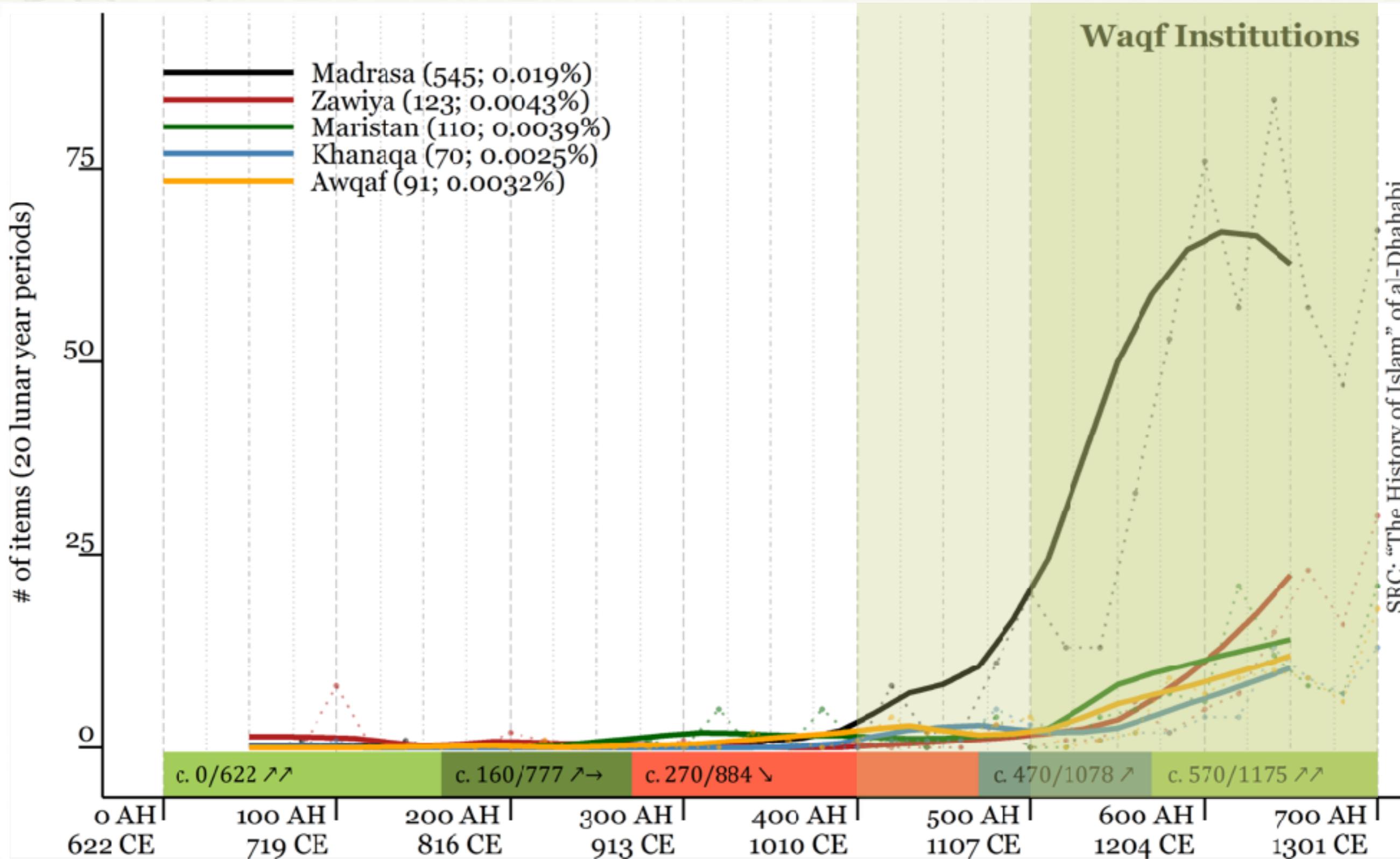
Professionalization and Institutionalization of the Learned



Professionalization and Institutionalization of the Learned



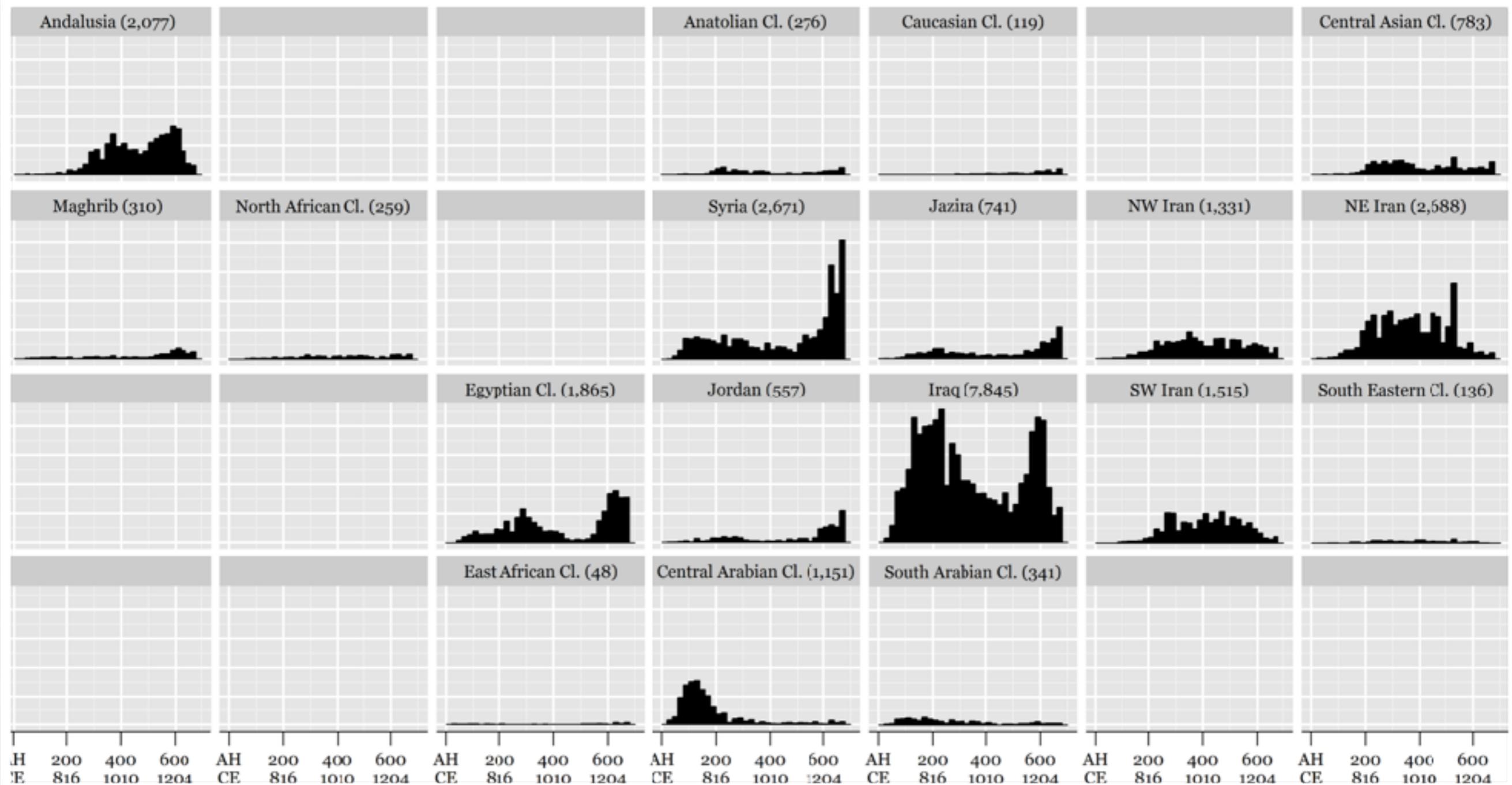
Professionalization and Institutionalization of the Learned



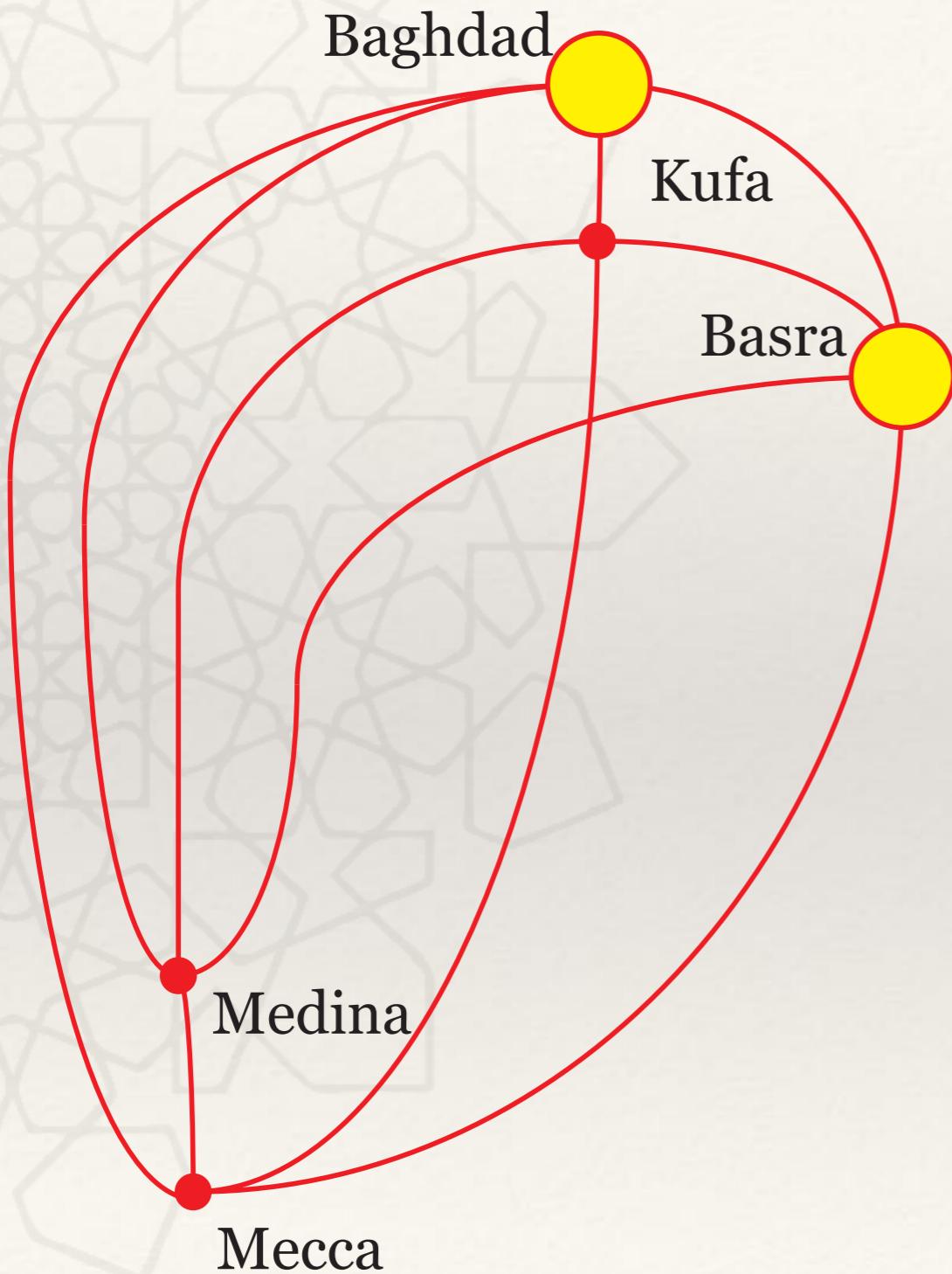


Social Geography

Geo-Chronological Coverage



Modeling Geographical Connections



- ❖ A cartogram of a person—al-Bağdādī and al-Basrī—whose biography mentions Baghdad, Kufa, Basra, Medina and Mecca
- ❖ *Such data can be grouped to show particular groups and/or periods*



The Core

anno domini
0632 - 0680 CE
Islamic World Changing

SRC: al-Dhahabi's *Ta'rikh al-islam*

Unadjusted dates: 40-90AH/661-709CE

380 760 1140 1530 1910 2290

Andalusia

Maghrib

North African Cluster

Anatolian Cluster

Jazira

Syria

Jordan

Iraq

Central Arabian Cluster

South Arabian Cluster

East African Cluster

NW Iran

NE Iran

SW Iran

South Eastern Cluster

anno hegirae
010 - 060

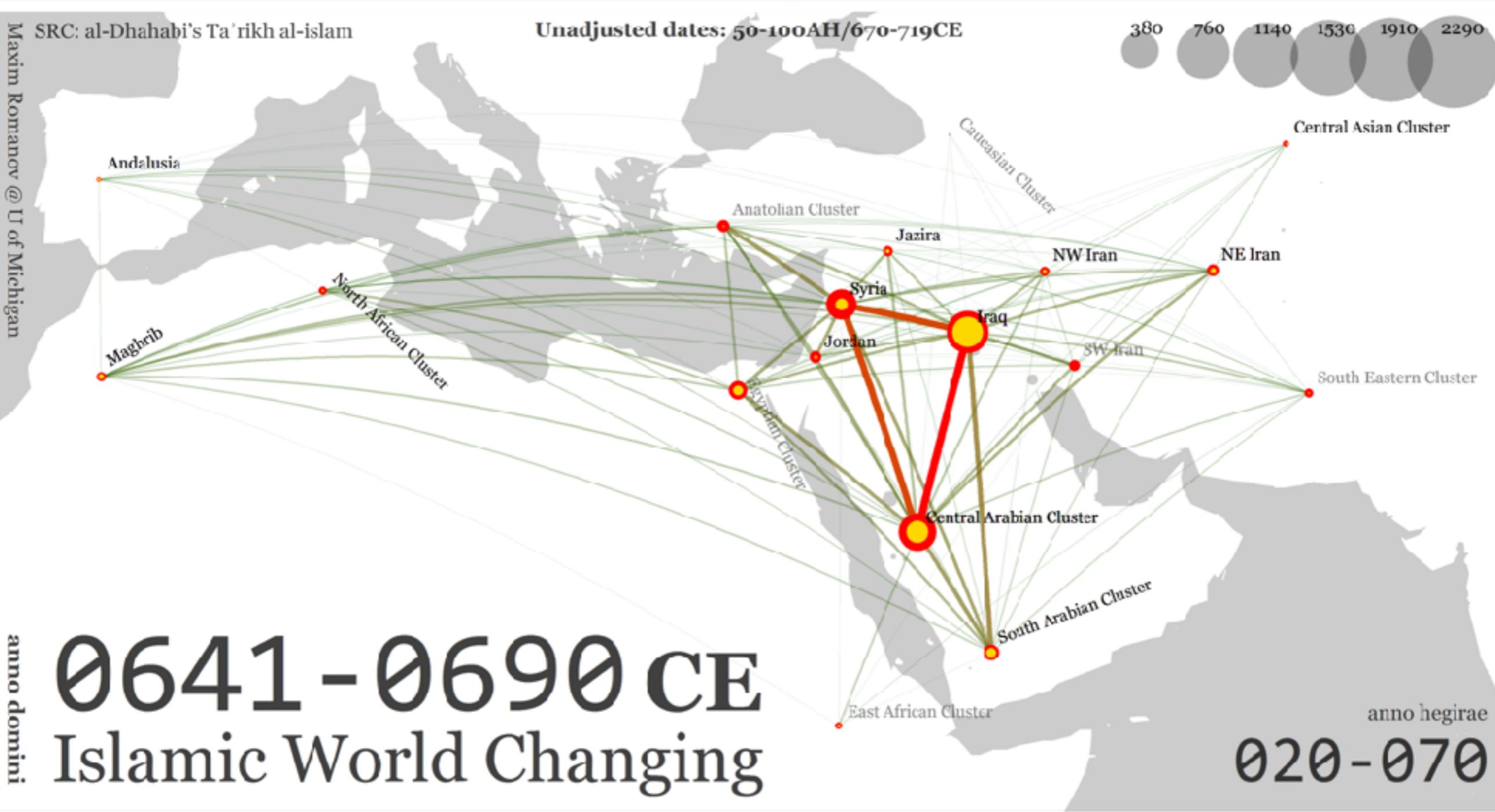
Caucasian Cluster

Central Asian Cluster

0641 - 0690 CE
Islamic World Changing

SRC: al-Dhahabi's *Ta'rikh al-islam*

Unadjusted dates: 50-100AH/670-719CE



anno hegirae

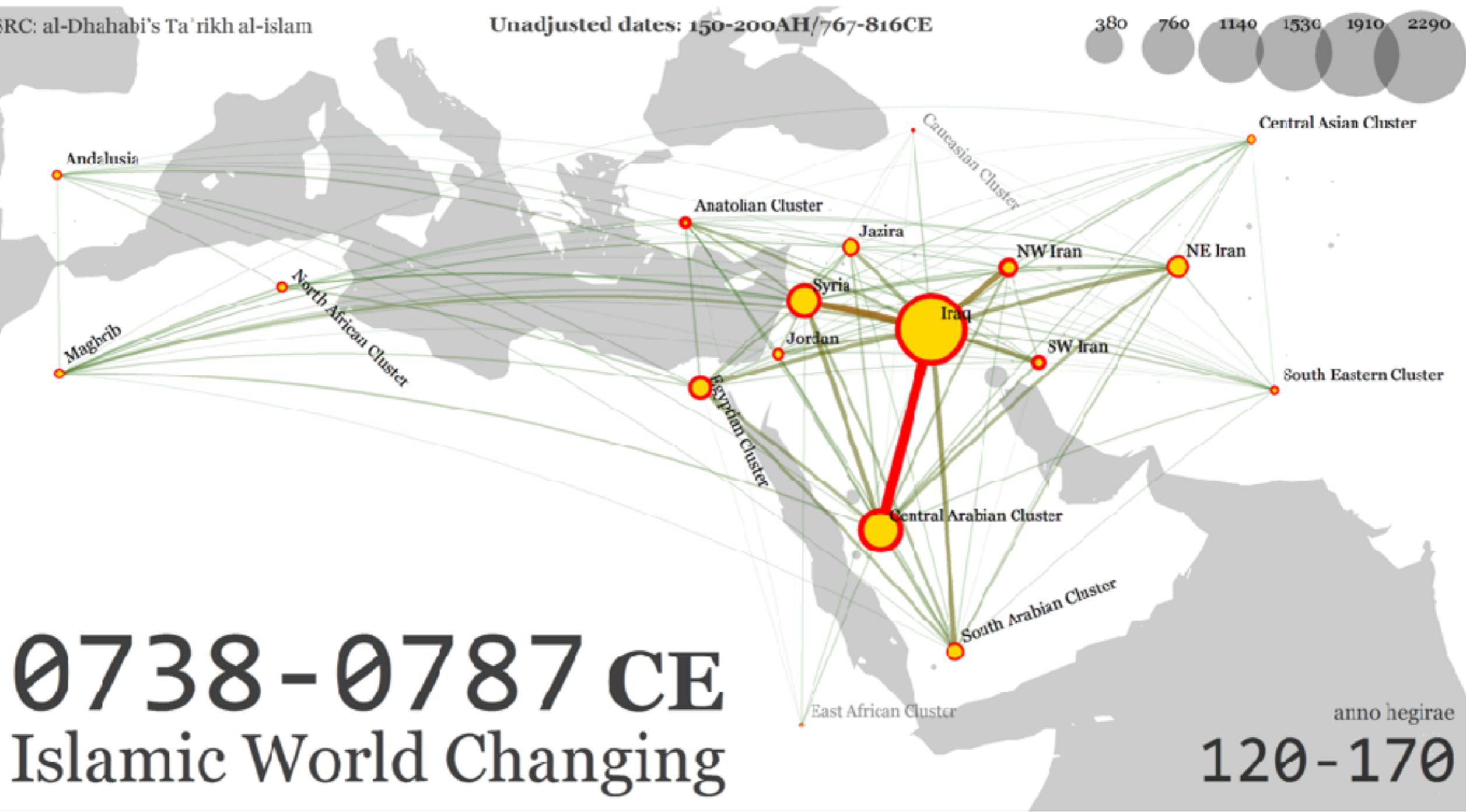
020 - 070

0738-0787 CE Islamic World Changing

SRC: al-Dhahabi's *Ta'rikh al-islam*

Unadjusted dates: 150-200AH/767-816CE

380 760 1140 1536 1910 2290



anno hegirae

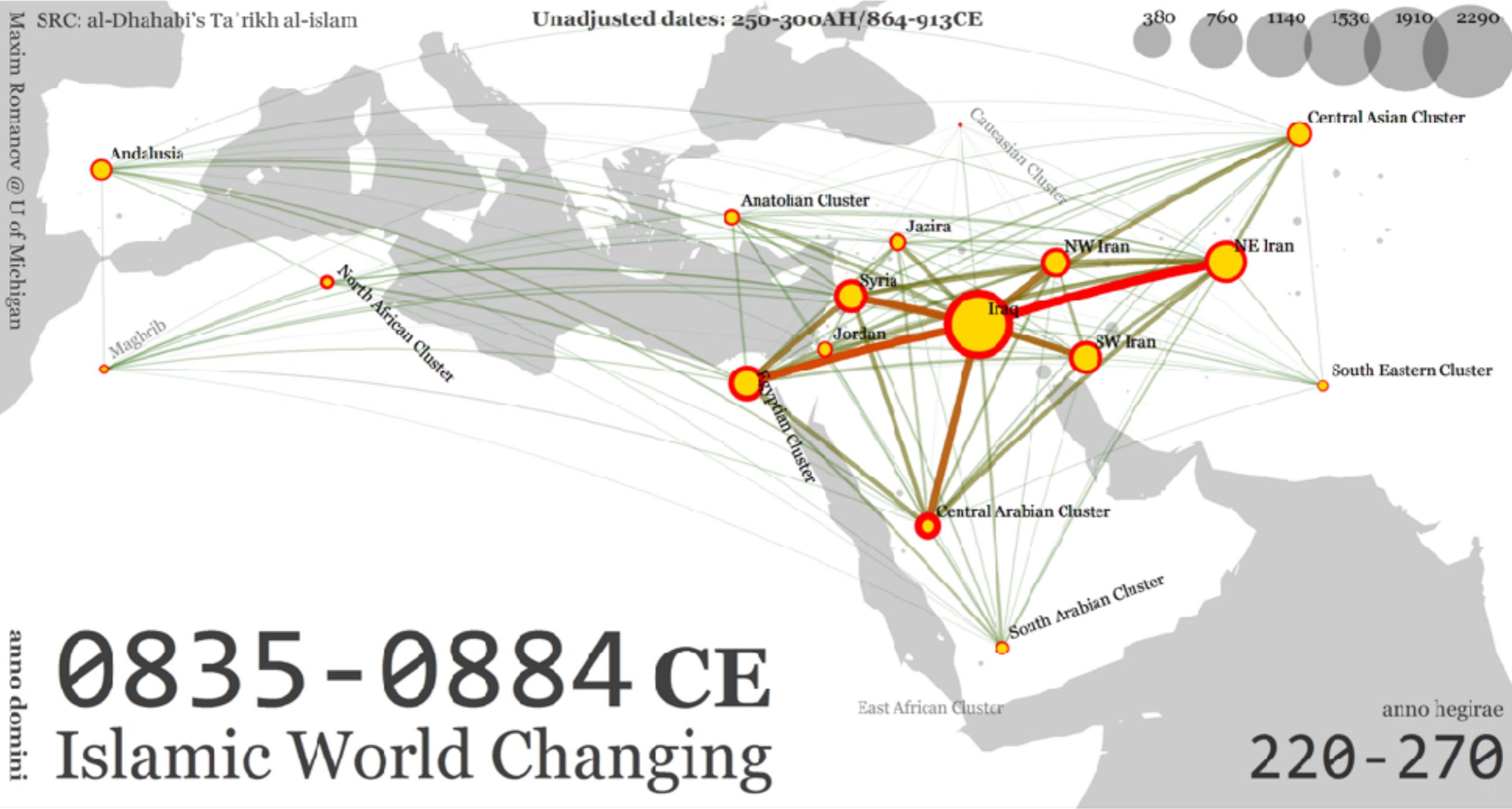
120-170

0835 - 0884 CE Islamic World Changing

SRC: al-Dhahabi's *Ta'rikh al-islam*

Unadjusted dates: 250-300AH/864-913CE

380 760 1140 1536 1910 2290

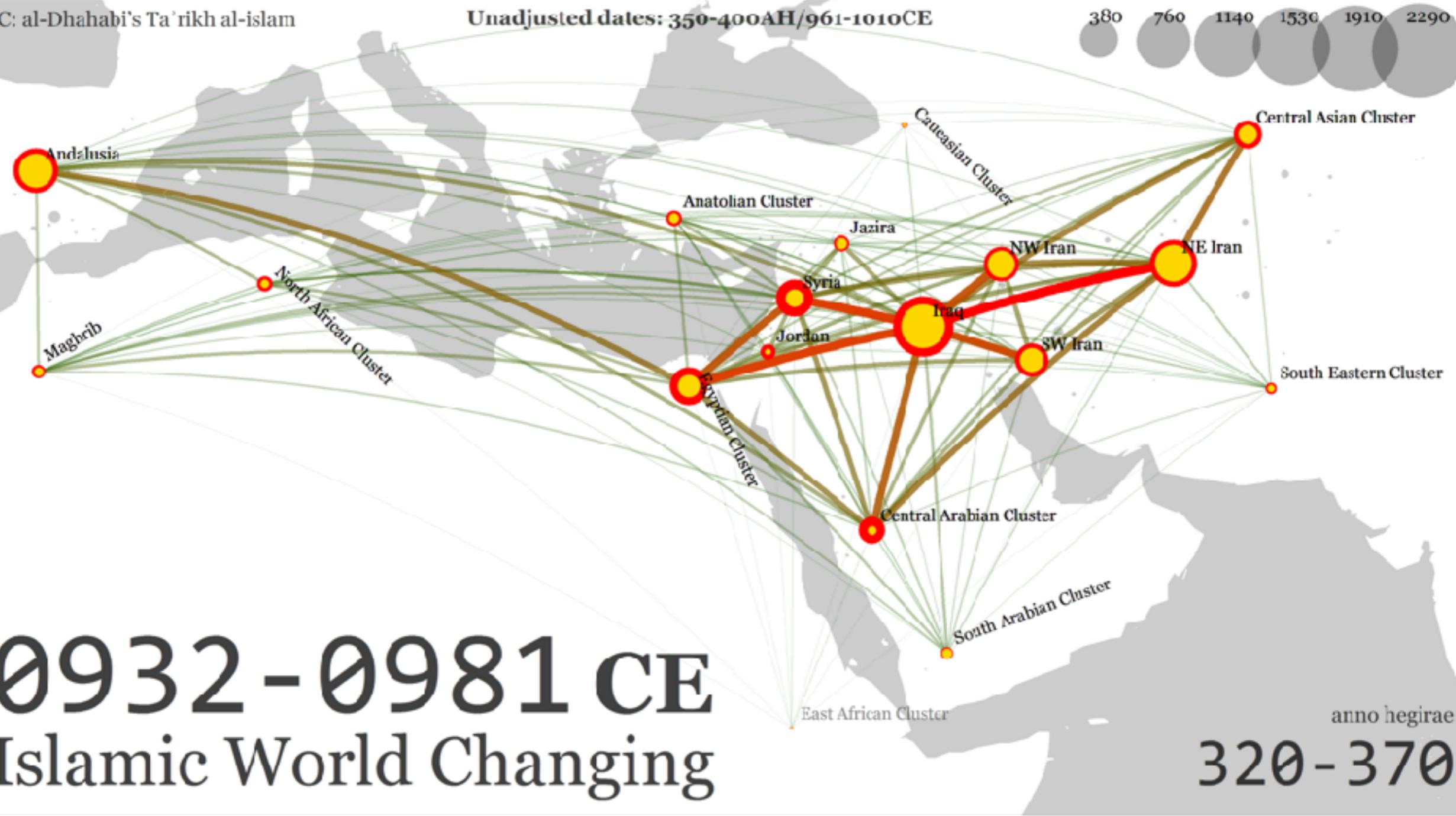


0932-0981 CE Islamic World Changing

SRC: al-Dhahabi's *Ta'rikh al-islam*

Unadjusted dates: 350-400AH/961-1010CE

380 760 1140 1536 1910 2290



anno hegirae

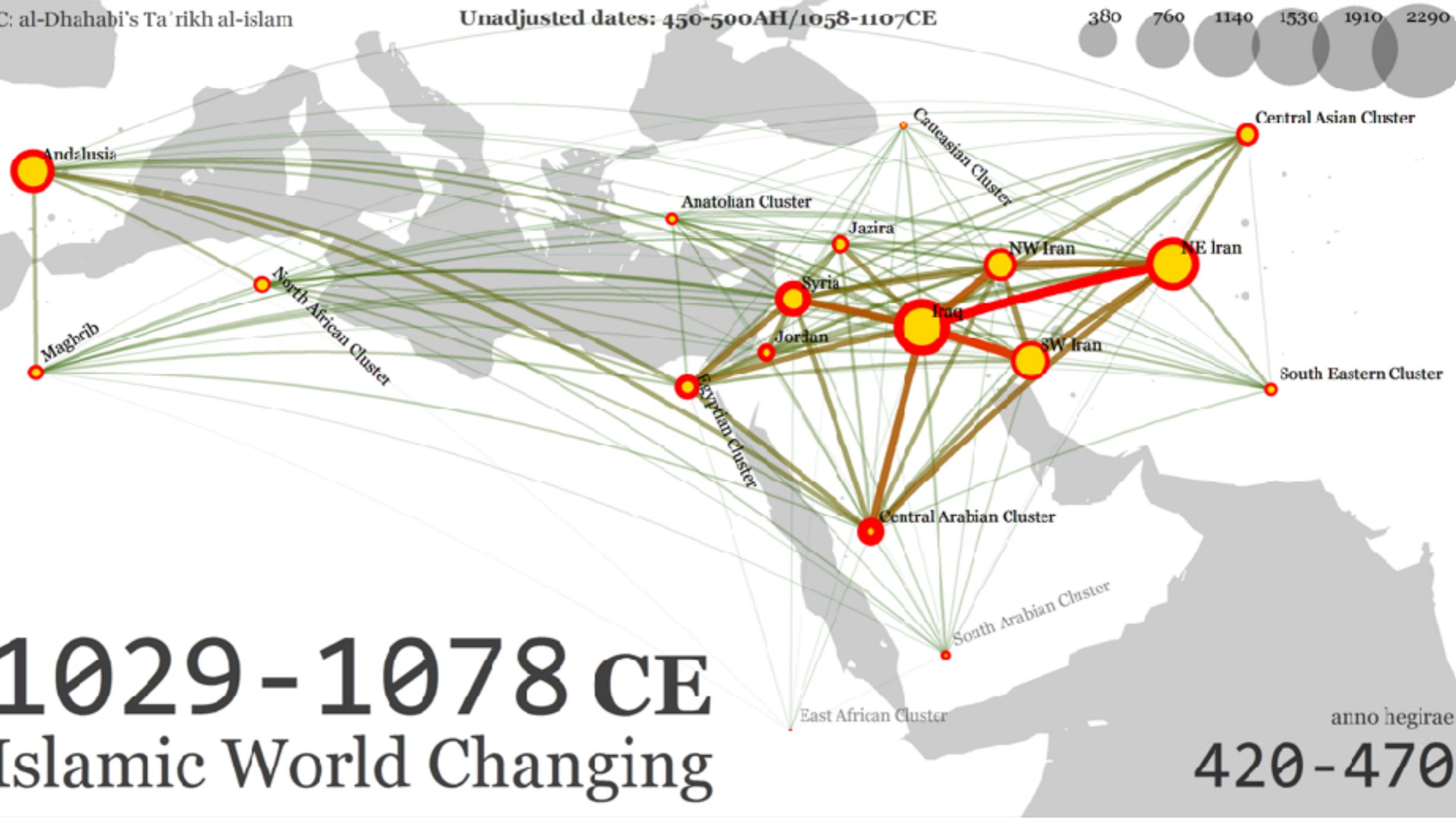
320-370

1029-1078 CE Islamic World Changing

SRC: al-Dhahabi's *Ta'rikh al-islam*

Unadjusted dates: 450-500AH/1058-1107CE

380 760 1140 1536 1910 2290



anno hegirae

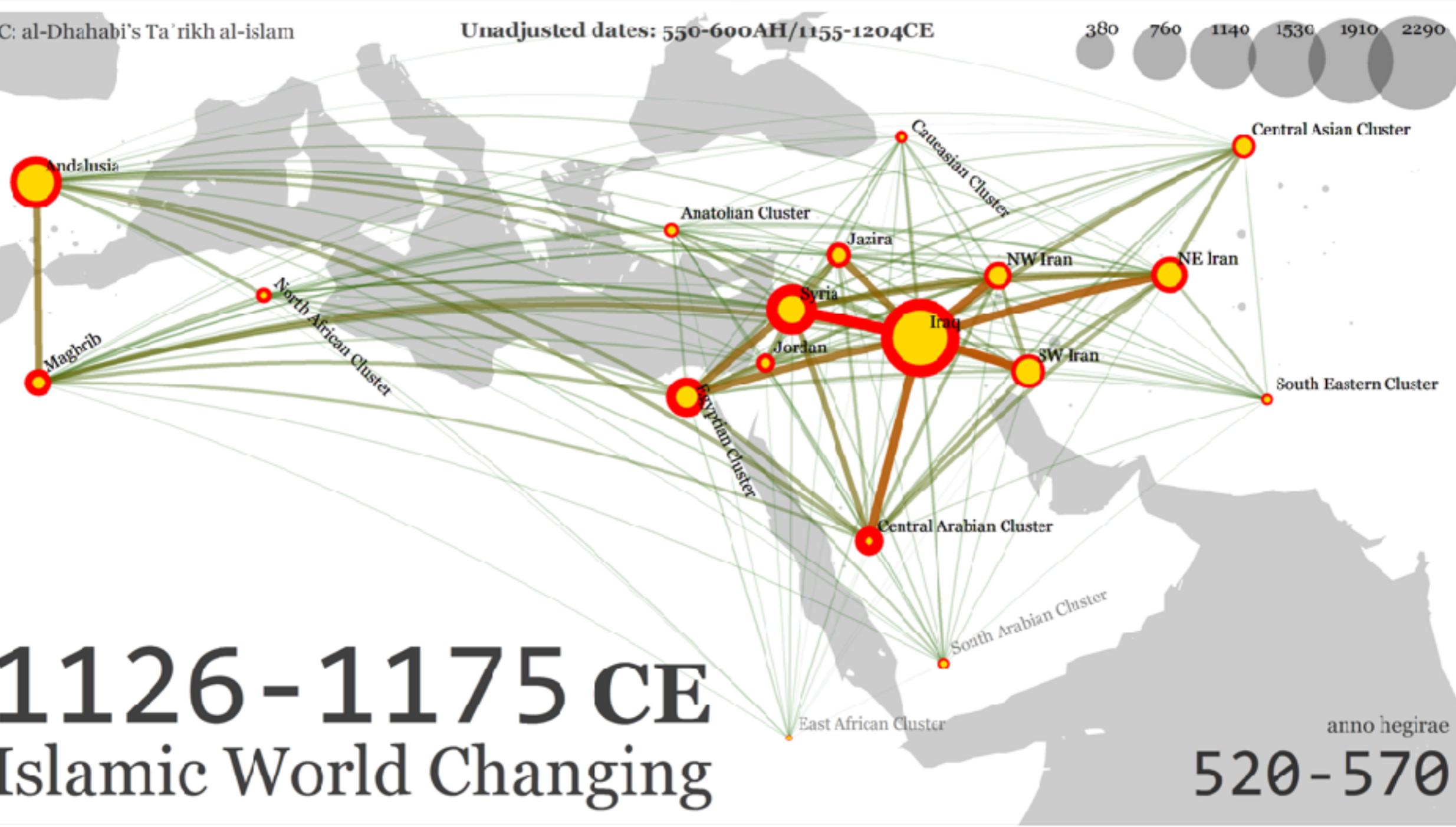
420-470

1126-1175 CE Islamic World Changing

SRC: al-Dhahabi's *Ta'rikh al-islam*

Unadjusted dates: 550-600AH/1155-1204CE

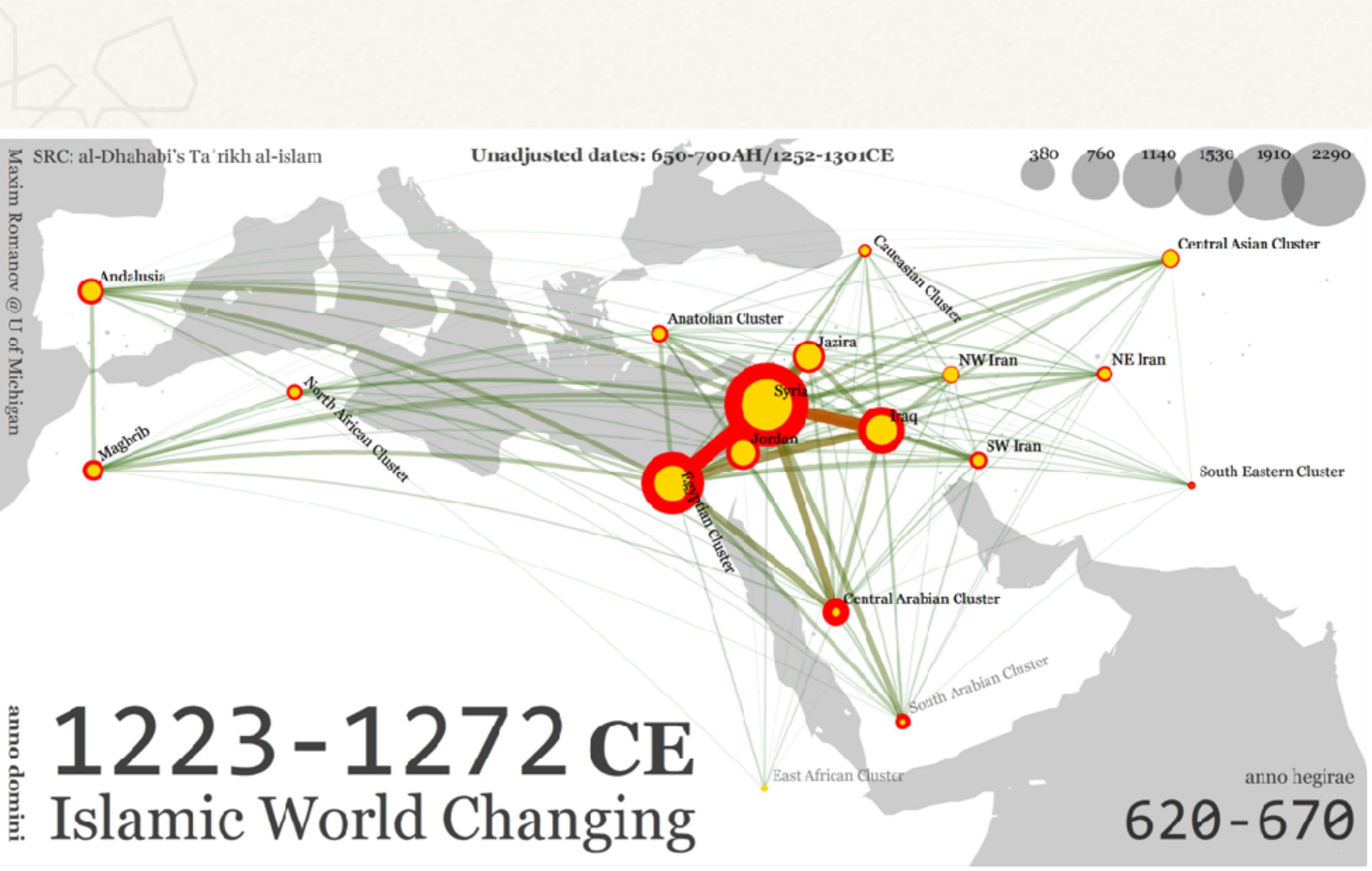
380 760 1140 1536 1910 2290

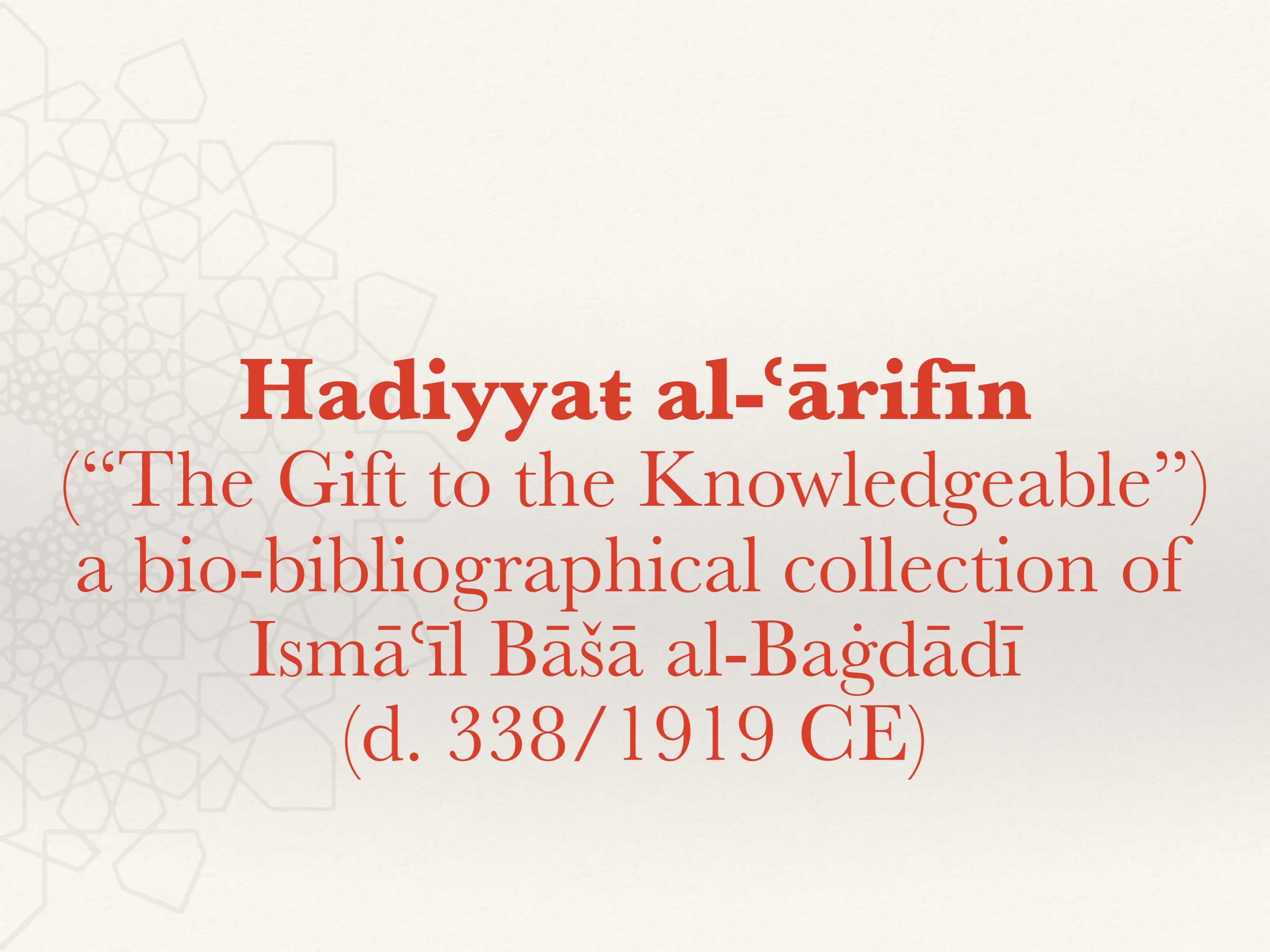


1223-1272 CE Islamic World Changing

SRC: al-Dhahabi's *Ta'rikh al-islam*

Unadjusted dates: 650-700AH/1252-1301CE





Hadiyyat al-‘ārifīn

(“The Gift to the Knowledgeable”)

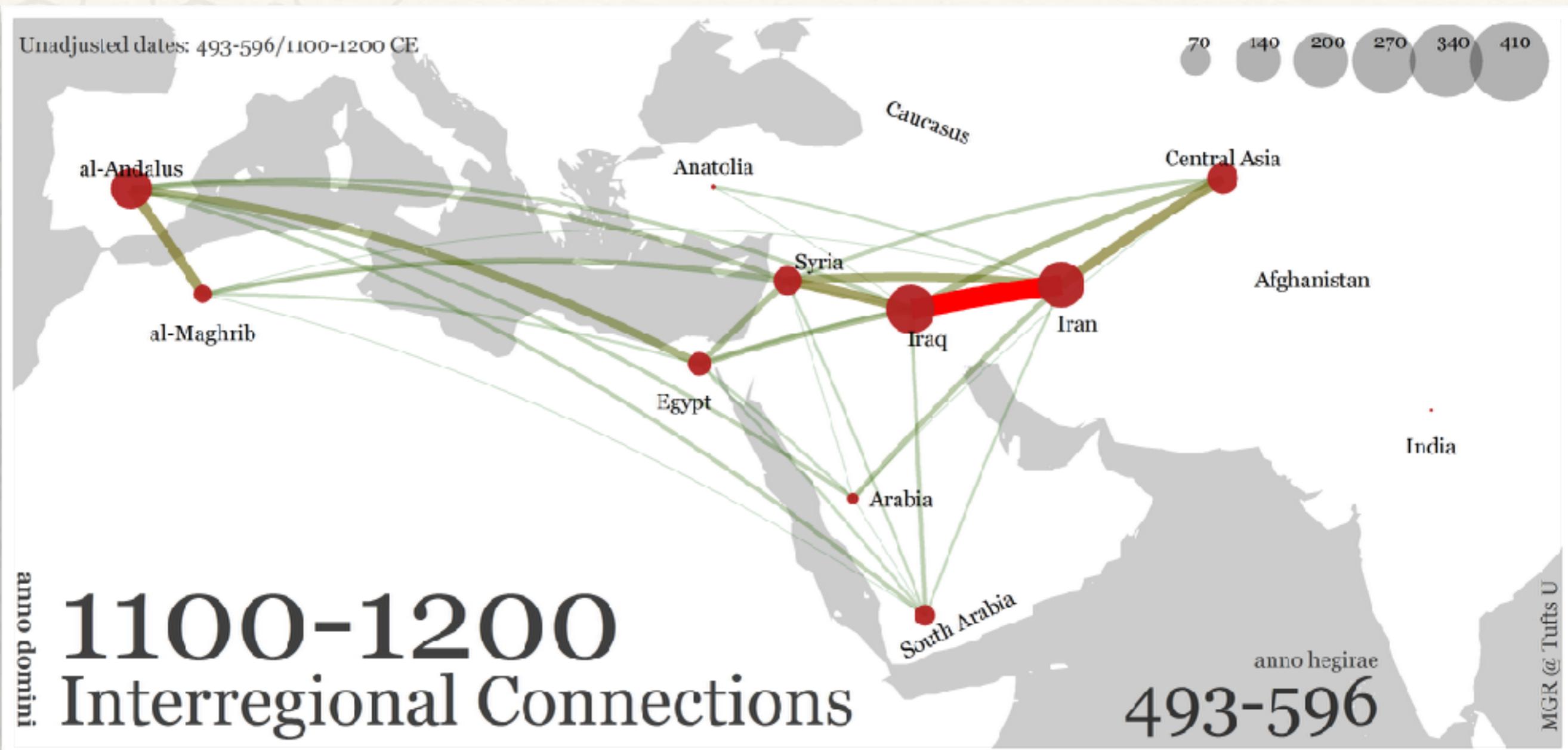
a bio-bibliographical collection of

Ismā‘il Bāšā al-Bağdādī

(d. 338/1919 CE)

Hadiyyat al-‘ārifīn (“The Gift to the Knowledgeable”)

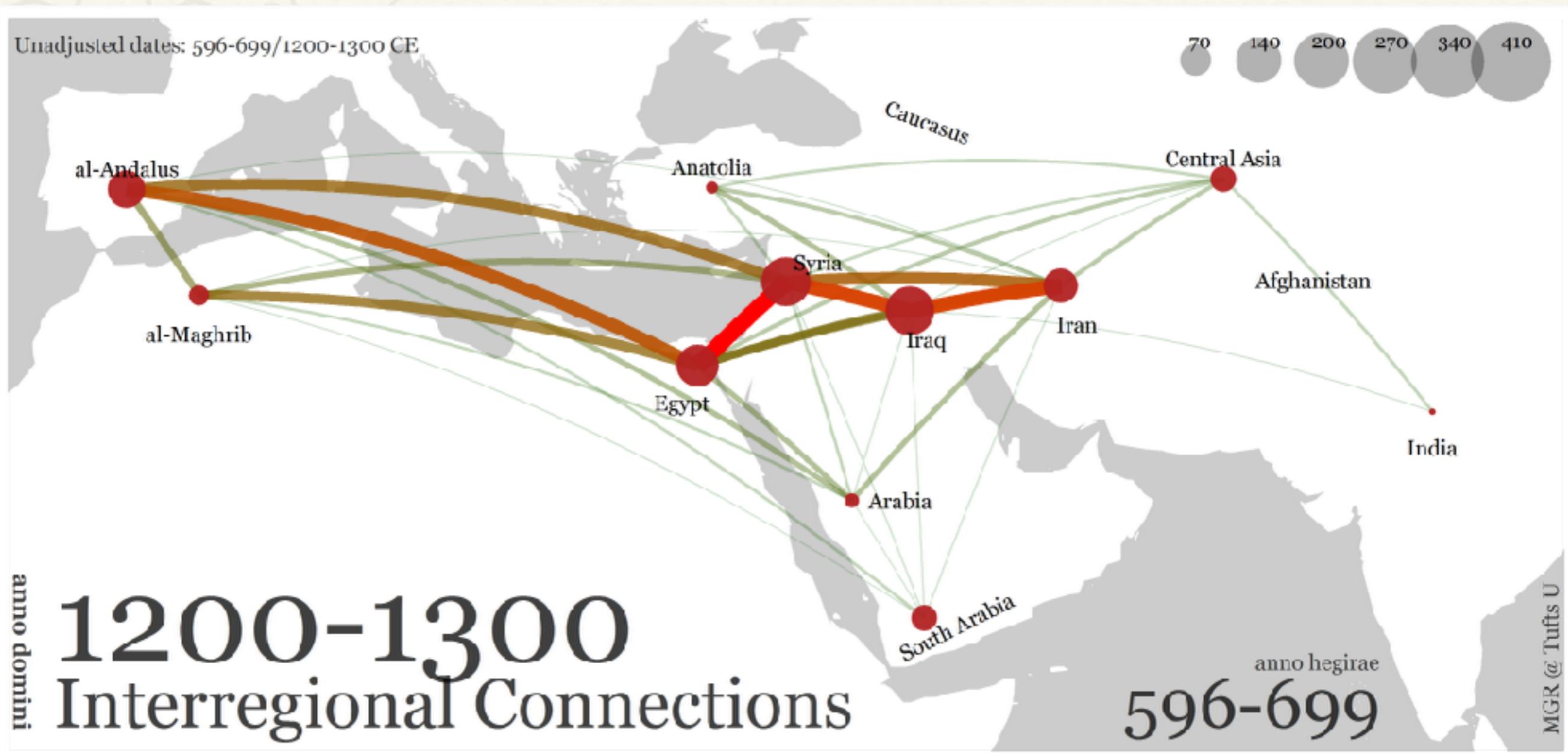
a bio-bibliographical collection of Ismā‘il Bāšā al-Bagdādī (d. 338/1919 CE)



The Iraqi-Iranian core up until the 12th century CE

Hadiyyat al-‘ārifīn (“The Gift to the Knowledgeable”)

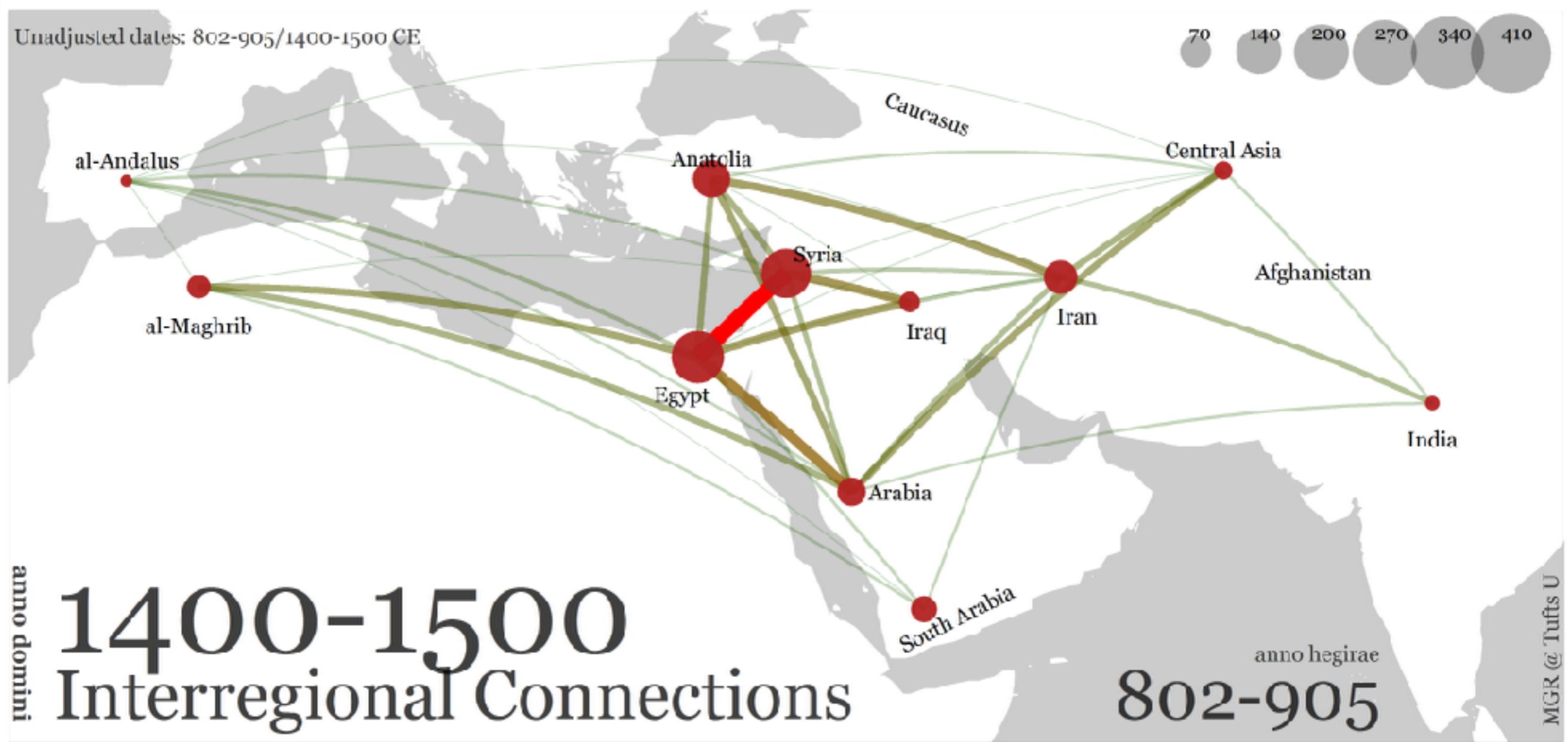
a bio-bibliographical collection of Ismā‘il Bāšā al-Bagdādī (d. 338/1919 CE)



Massive Migrations in the 13th century CE

Hadiyyat al-‘ārifīn (“The Gift to the Knowledgeable”)

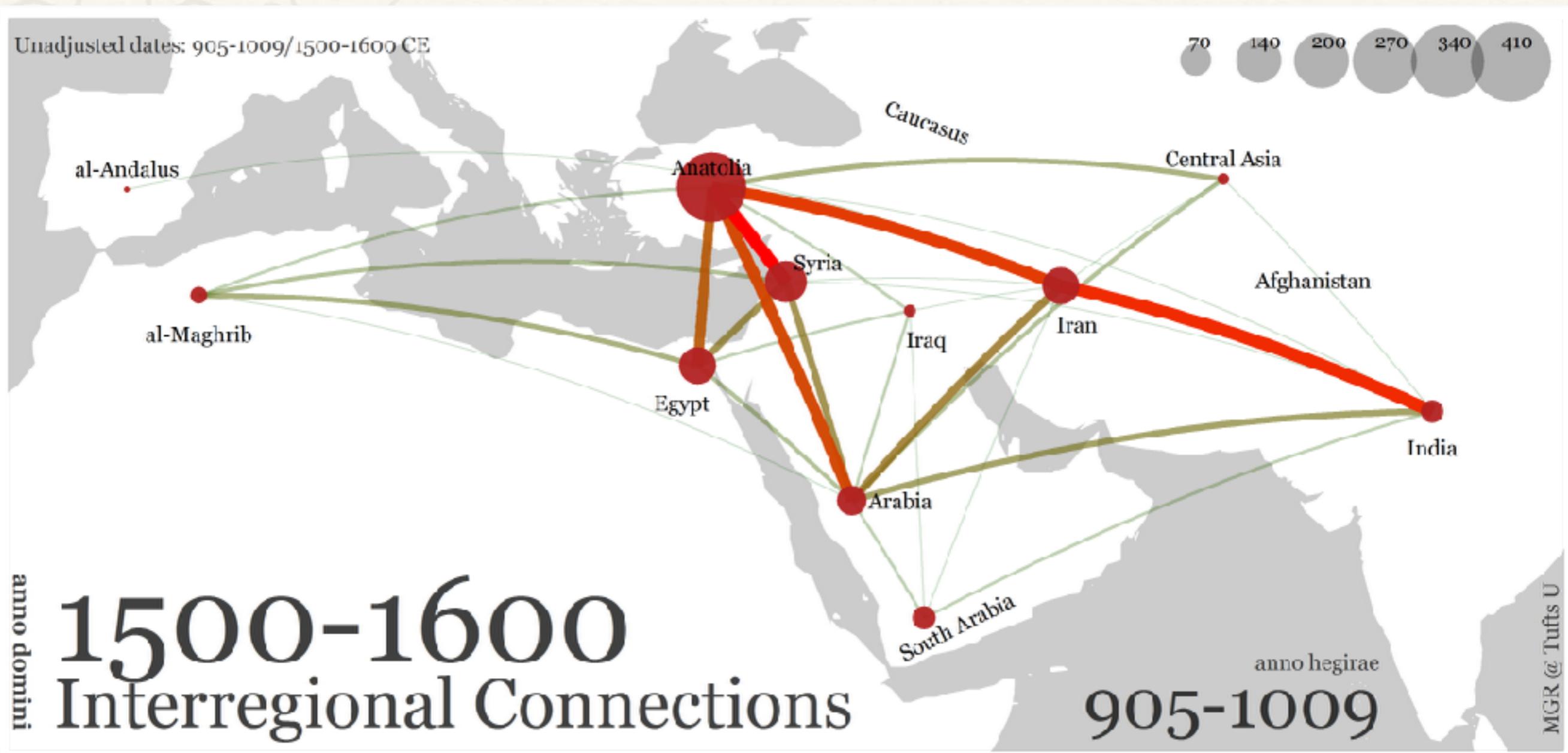
a bio-bibliographical collection of *Ismā‘il Bāšā al-Bagdādī* (d. 338/1919 CE)



The New Mamlūk Core in the 14th and 15th Centuries CE

Hadiyyat al-‘ārifīn (“The Gift to the Knowledgeable”)

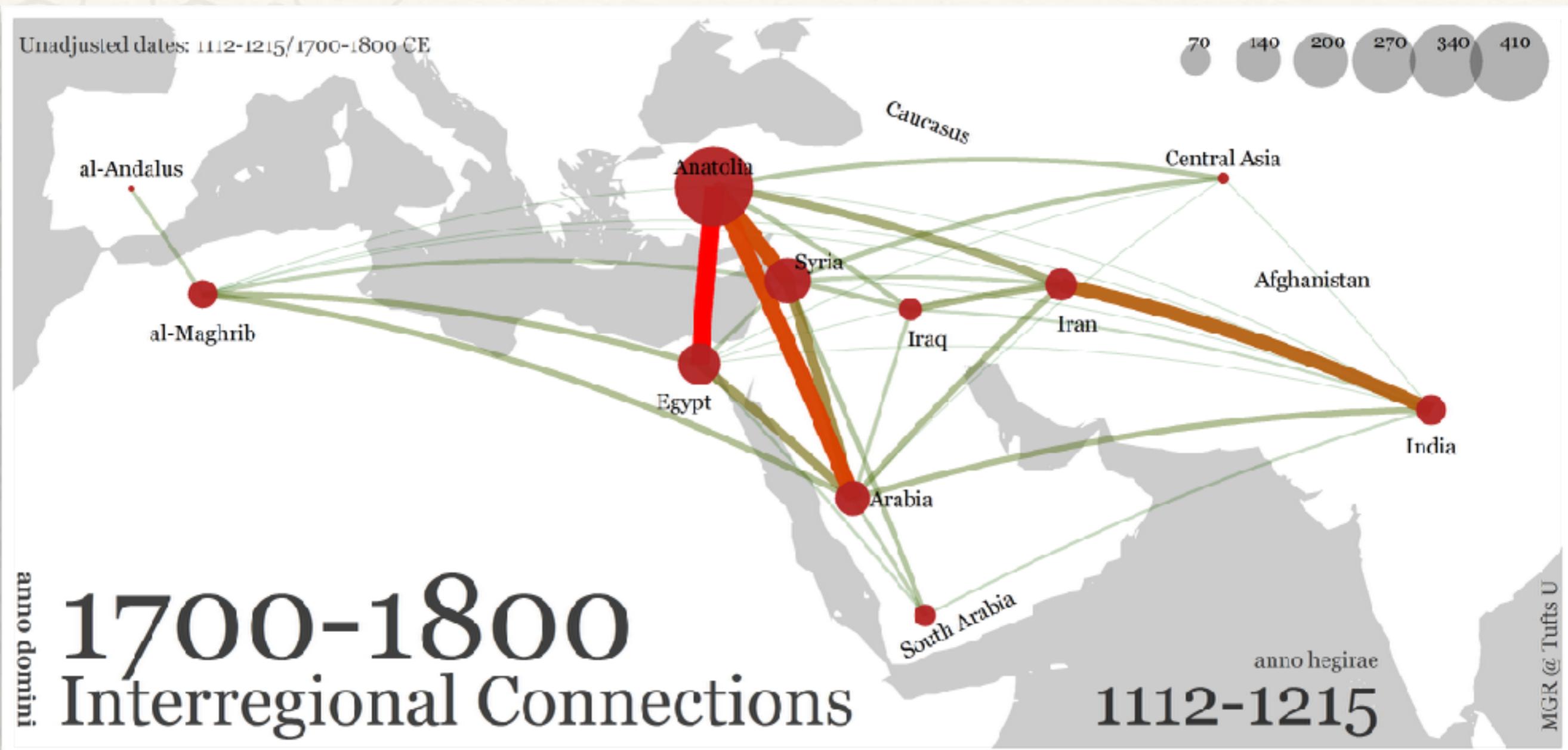
a bio-bibliographical collection of Ismā‘il Bāšā al-Bagdādī (d. 338/1919 CE)



Reconfiguration of the 16th Century CE

Hadiyyat al-‘ārifīn (“The Gift to the Knowledgeable”)

a bio-bibliographical collection of *Ismā‘il Bāšā al-Bagdādī* (d. 338/1919 CE)



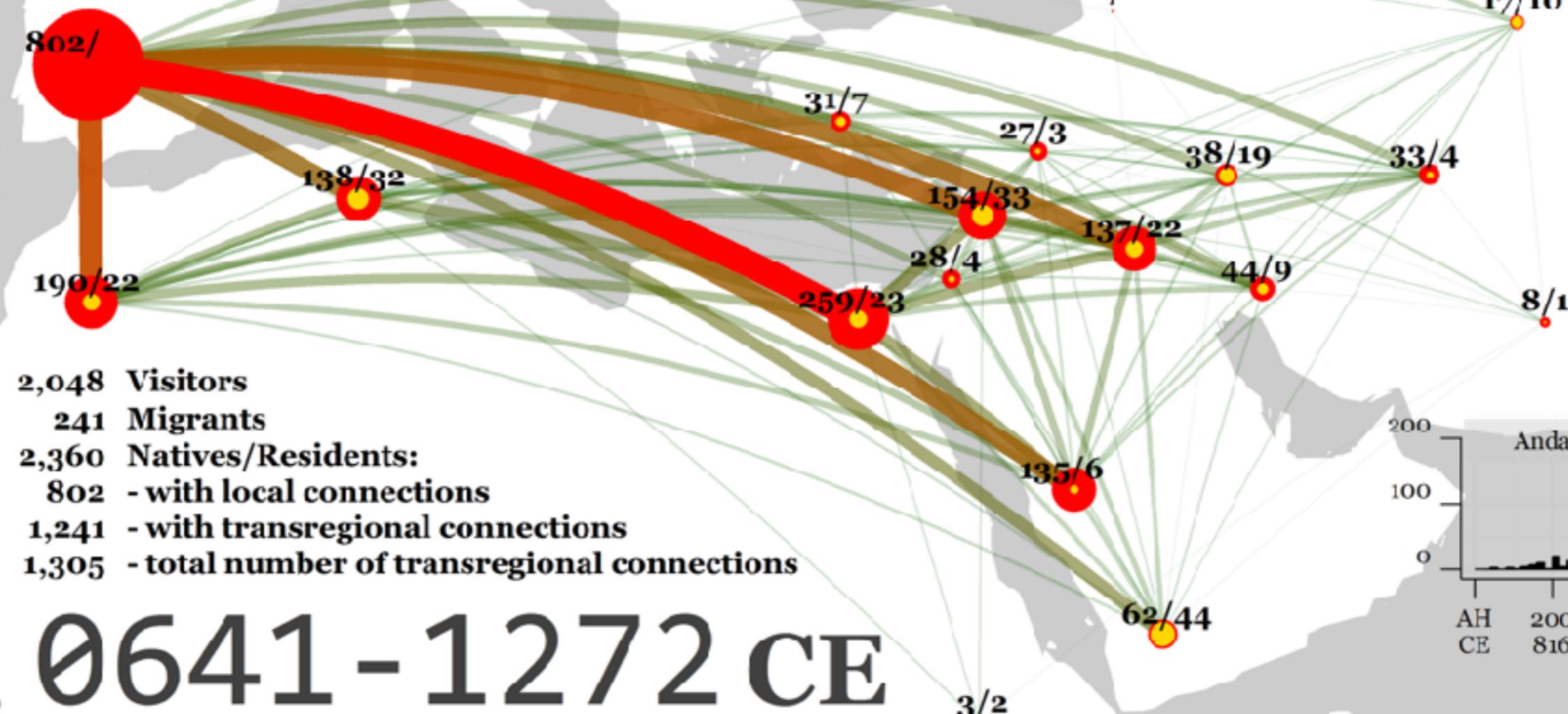
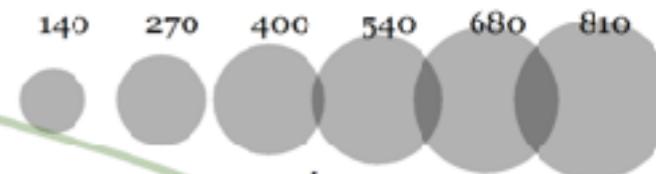
The Turco-Arabic and Indo-Iranian Cores:
configuration of 17th–19th centuries



**Regional Integration
Social and religious groups
Urban centers**

SRC: al-Dhahabi's Ta'rikh al-islam

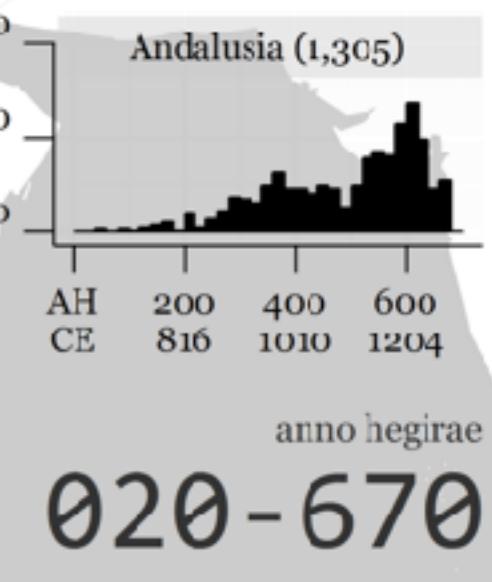
Unadjusted dates: 50-700AH/670-1301CE



anno domini:

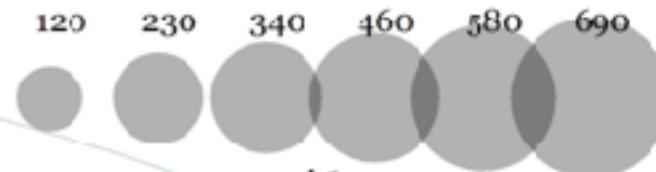
0641-1272 CE

Connections: Andalusia



SRC: al-Dhahabi's Ta'rikh al-islam

Unadjusted dates: 50-700AH/670-1301CE



14/4

8/3

47/5

2,312 Visitors**526** Migrants**2,736** Natives/Residents: **452** - with local connections **1,398** - with transregional connections **2,077** - total number of transregional connections

38/7

143/15

39/10

249/81

54/10

684/192

20/2

181/42

138/42

452/

49/13

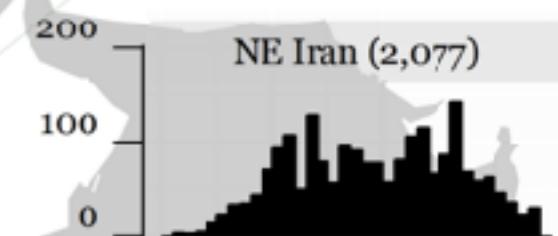
214/30

47/2

0641-1272 CE

Connections: NE Iran

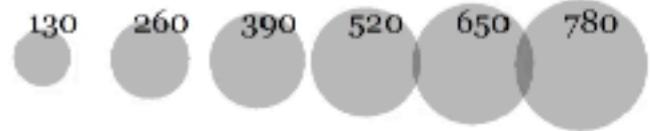
NE Iran (2,077)

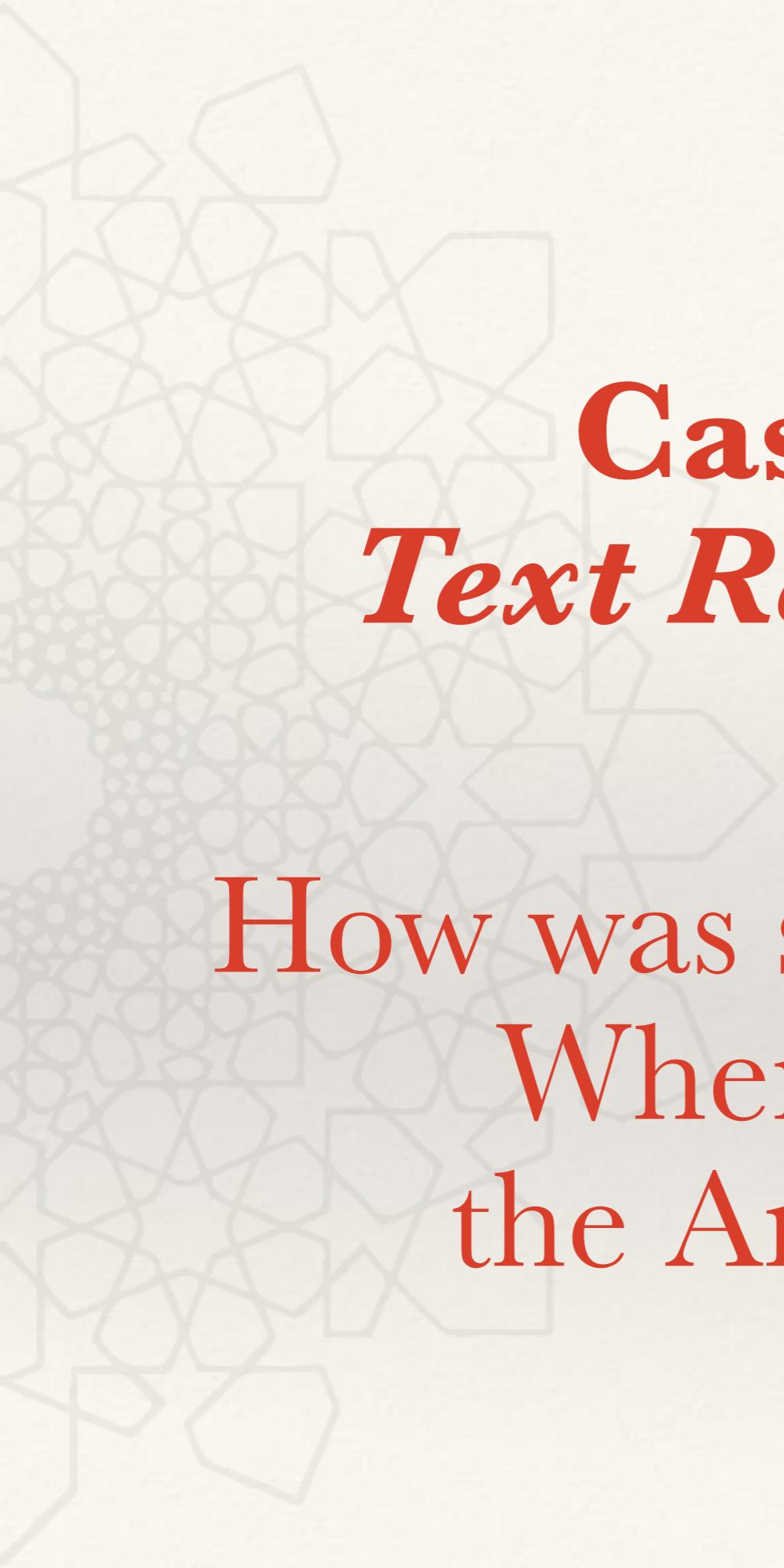
**020-670**



SRC: al-Dhahabi's *Ta'rikh al-islam*

Unadjusted dates: 40-70AH/661-690CE





Case Study II: *Text Reuse Detection*

How was such a text written?
Where does it sit in
the Arabic tradition?

“Compilation”?

- ❖ How does one write a book like that w/o copy-paste?!
 - ❖ Did he paraphrase his sources?
 - ❖ changing their language, and thus bringing 14th century language into the descriptions of earlier periods
 - ❖ Did he “stitch” the TI from his sources?
 - ❖ thus preserving the language of earlier sources

KITAB Project: kitab-project.org



Sarah Savant, Associate Professor, AKU-ISMC, and Principle Investigator of KITAB. Dr. Savant is a cultural historian specialising in the Middle East and Iran ca. 600-1100. Her work on KITAB is motivated by a desire to write a cultural history of the Arabic book using digital tools. Her publications include *The New Muslims of Post-Conquest Iran: Tradition, Memory, and Conversion* (Cambridge University Press, 2013; winner of the Said-Sirjani book award given by the International Society for Iranian Studies on behalf of the Persian Heritage Foundation).



Ahmad Sakhi is an IT professional, specializing in Enterprise Information and Content Management space. He is currently consulting for Capgemini, a global consulting firm. Mr. Sakhi has a degree in Computer Information Systems and has completed various projects in the Finance, Telecom, Banking and Energy sectors in North America, including designing, building architecture and implementing end-to-end solutions. Mr. Sakhi is from Toronto, Canada and is involved in the KITAB project as a TKN volunteer.



Malik Merchant is a Software Consultant, specializing in Microsoft Technologies. He currently is working for Avanade, a joint venture between global consulting firm, Accenture and Microsoft. Mr. Merchant has a Bachelor's degree in Computer Science from Mumbai, India and has been involved in several projects across several industries with focus in Oil & Gas and Insurance domain. He now lives in Calgary, Canada and is involved in the KITAB project as a TKN volunteer.



Sohail Merchant is currently working as the Assistant Registrar for the Aga Khan University. Prior to working at AKU-ISMC, he has worked with various organisations like City Hampers, Child-to-Child – University of London and Blue Systems Inc. where he has designed, managed and implemented various solutions and web applications. Sohail is a Microsoft Certified Professional and holds a Master of Science degree in Computing from London Metropolitan University.

KITAB Project: kitab-project.org



Gregory Ralph Crane

Alexander von Humboldt Professor of Digital Humanities



Digital Humanities
UNIVERSITÄT LEIPZIG

Greg completed his doctorate in classical philology at Harvard University and subsequently worked there as an assistant professor. From 1985, he was involved in planning the Perseus Project as a co-director and is now its Editor-in-Chief. He was associate professor at TUFTS University. He has received, among others, the Google Digital Humanities Award 2010 for his work in the field.



David Smith is an Assistant Professor in the College of Computer and Information Science at Northeastern University and a founding member of the NULab for Texts, Maps, and Networks, Northeastern's center for the digital humanities and computational social sciences. Previously, he was a research faculty member at the University of Massachusetts' Center for Intelligent Information Retrieval, a Ph.D. student in computer science at Johns Hopkins University, and the head programmer at the Perseus Digital Library Project. His research focuses on building statistical models of human language, with applications to information retrieval, machine translation, the humanities, and social sciences. Most recently, he has been working on inference for social networks from textual evidence, in collaboration with colleagues in English, history, and political science, under the aegis of the Proteus and Viral Texts projects.

<http://viraltexts.org/> Ryan Cordell, David Smith (Code), and others



How to address this issue?

Passage1: JK000982_000292

Passage2: Shamelah0023775_003346

Begin Position: 1

End Position: 58

Text reuse: computational methods of tracing long quotations, paraphrases, allusions, etc.

al-Ta‘ālibī’s *Timār al-qulūb fī-l-mudāf wa-l-mansūb* and one of its sources

Text1:

عند عبيد الله بن زياد اذ ادخل علي-----ه جرذا ابيض فتعجب منه ف قال --
لعبد الله----- يا ابو صالح هل رأيت اعجب من هذا-----
و اذا عبد الله قد تفاءل----- كانه فrex و اصف----ر كانه جرادة---
فقال عبيد الله ابو صالح يعصي الرحمن ويتهانون **بالسلطان** ويقبض على
الثعبان ويمشي الي الاسد الورد ويلقي الرماح بوجهه والسيوف بيده وقد
اعتراه من

Text2:

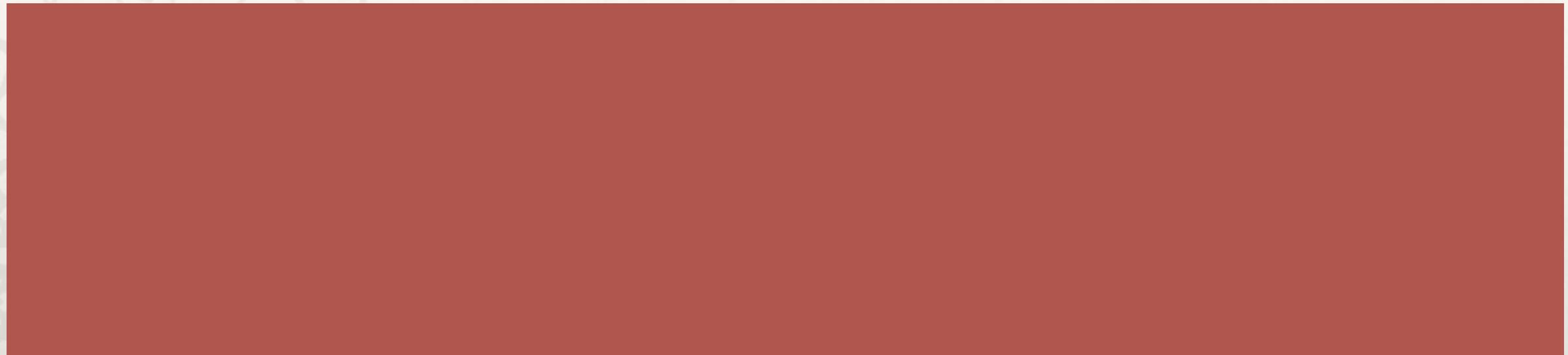
عند عبيد الله بن زياد اذ ادخل علي عبد الله جرذا - ابيض ليعجب منه فا قبل عبيد
الله علي عبد الله فقال هل رأيته يا ابو صالح----- اعجب من هذا الجرذا فقط
و اذا عبد الله قد تفاءل حتى صار كانه فrex و اصغر حتى صار كانه جرادة ذكر
فقال عبيد الله ابو صالح يعصي الرحمن ويتهانون **باليشيطان** ويقبض على الثعبان
ويمشي الي الاسد ويلقي الرماح بوجهه ----- وقد اعتراه من

Graphing Method



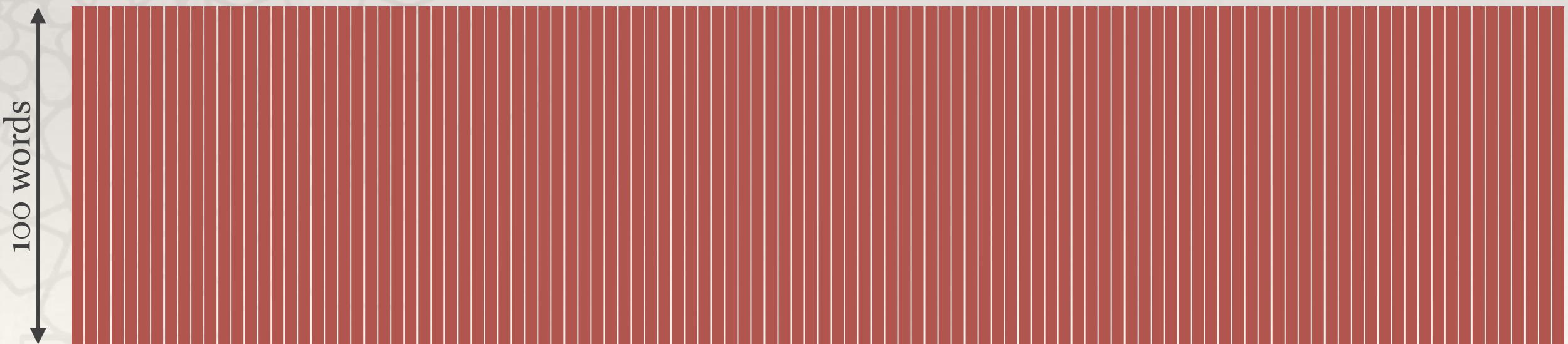
Book A

Graphing Method



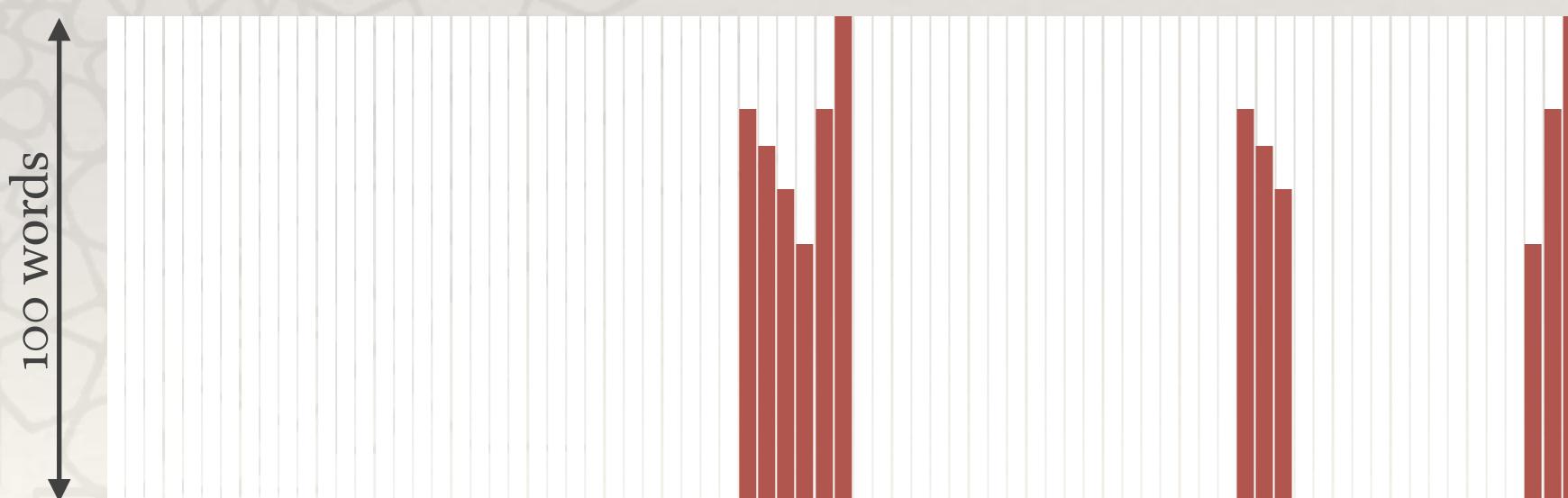
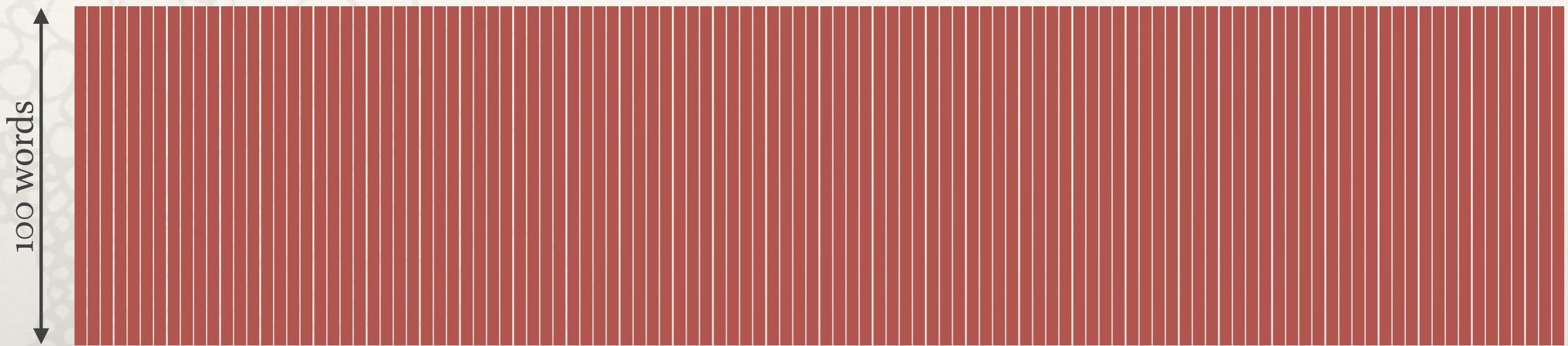
Beginning of the book

End of the book



Abstraction

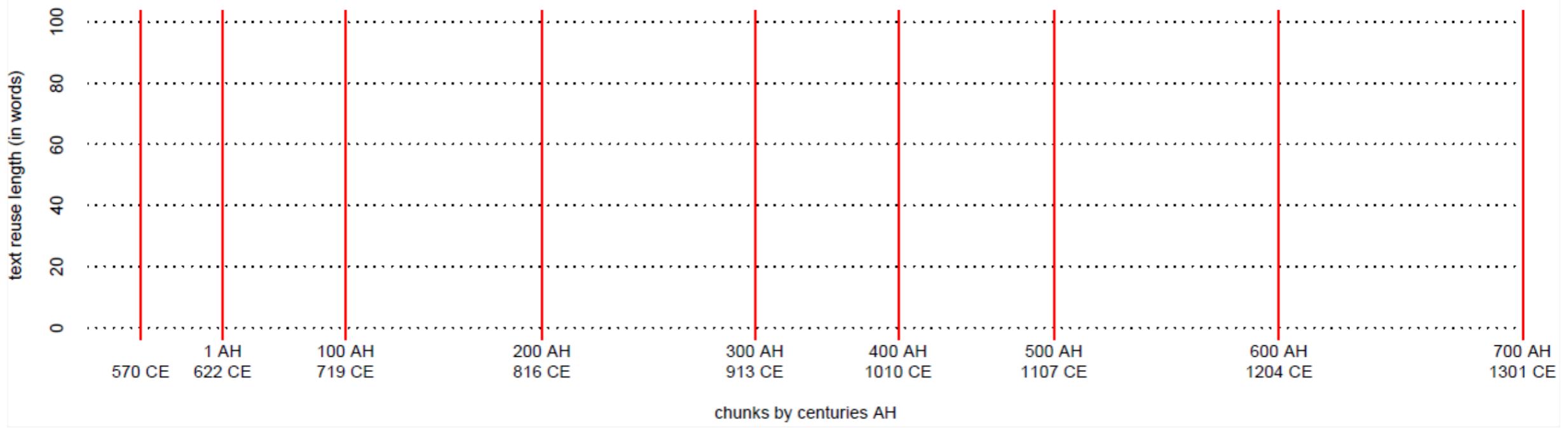
Graphing Method



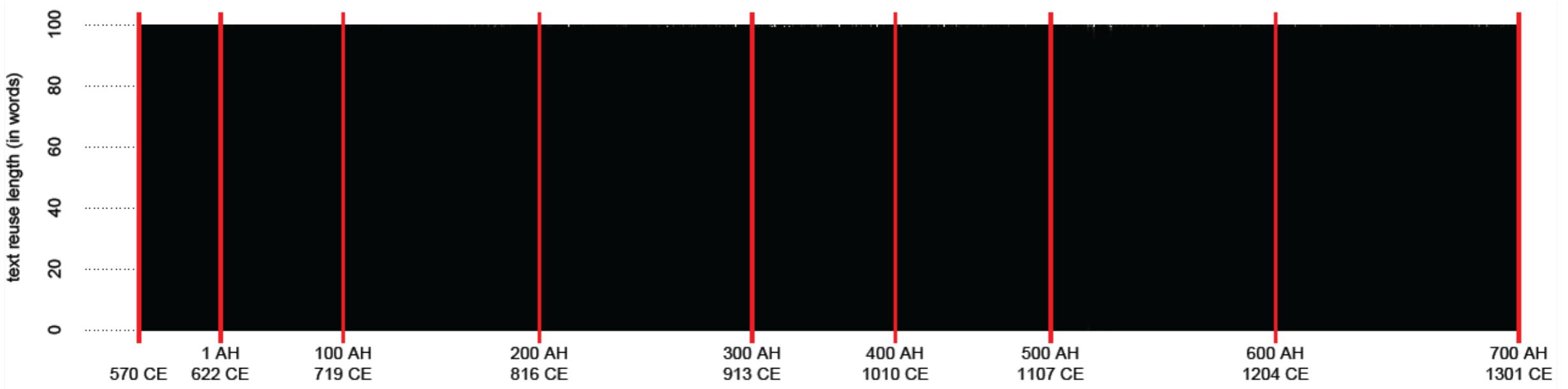
Abstraction

Text Reuse in the *Ta'rikh al-islam*

Ta'rikh al-islam chunked and divided into periods

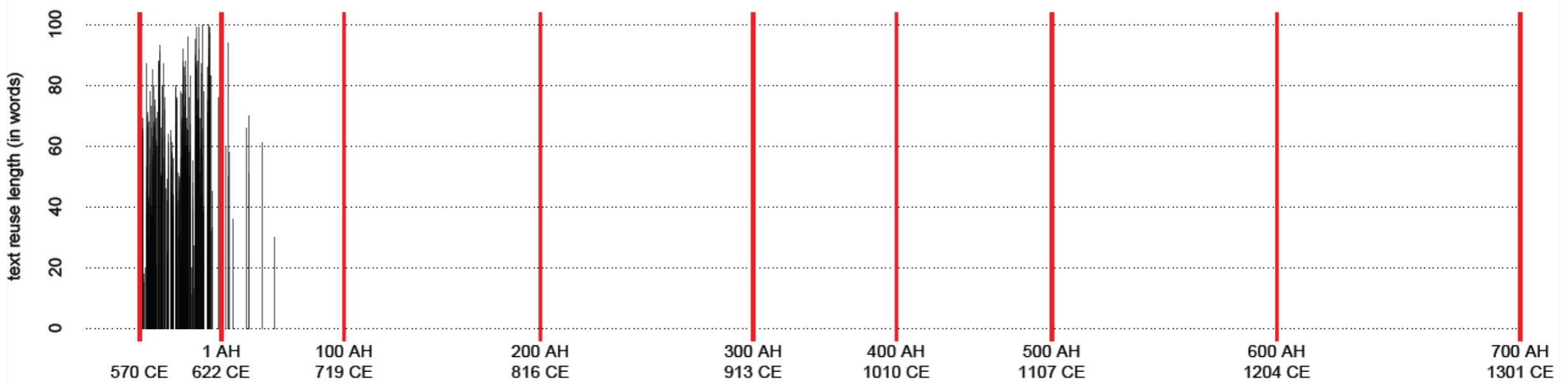


Ta'rikh al-islam compared with itself

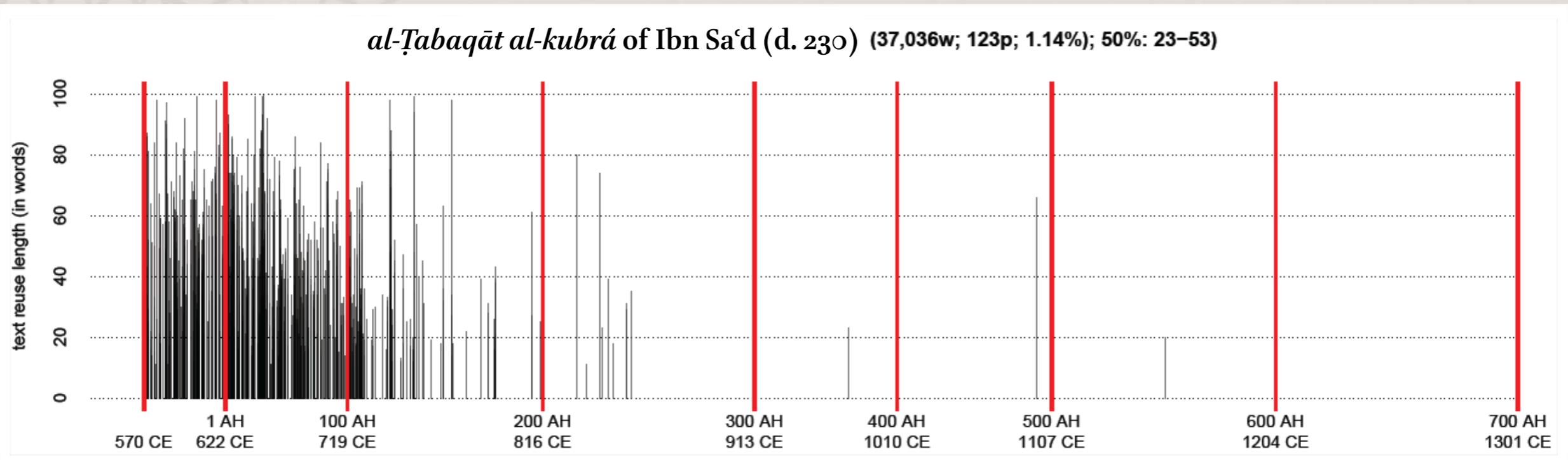


Text Reuse in the *Ta'rih al-islām*

al-Sīrat al-Nabawīyat of Ibn Hišām (d. 213) (28,339w; 94p; 0.872%); 50%: 30–63)

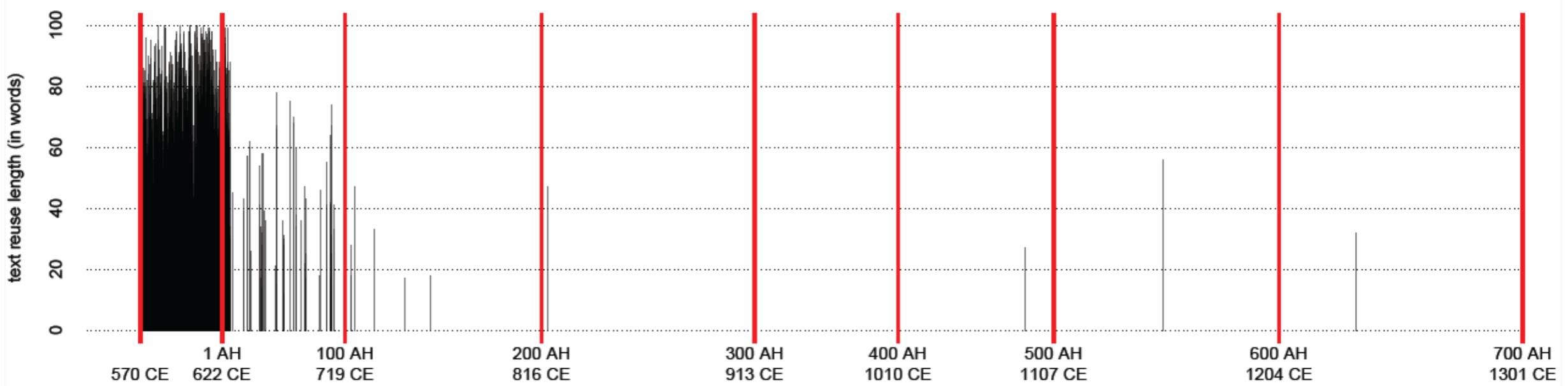


al-Tabaqāt al-kubrā of Ibn Sa'd (d. 230) (37,036w; 123p; 1.14%); 50%: 23–53)

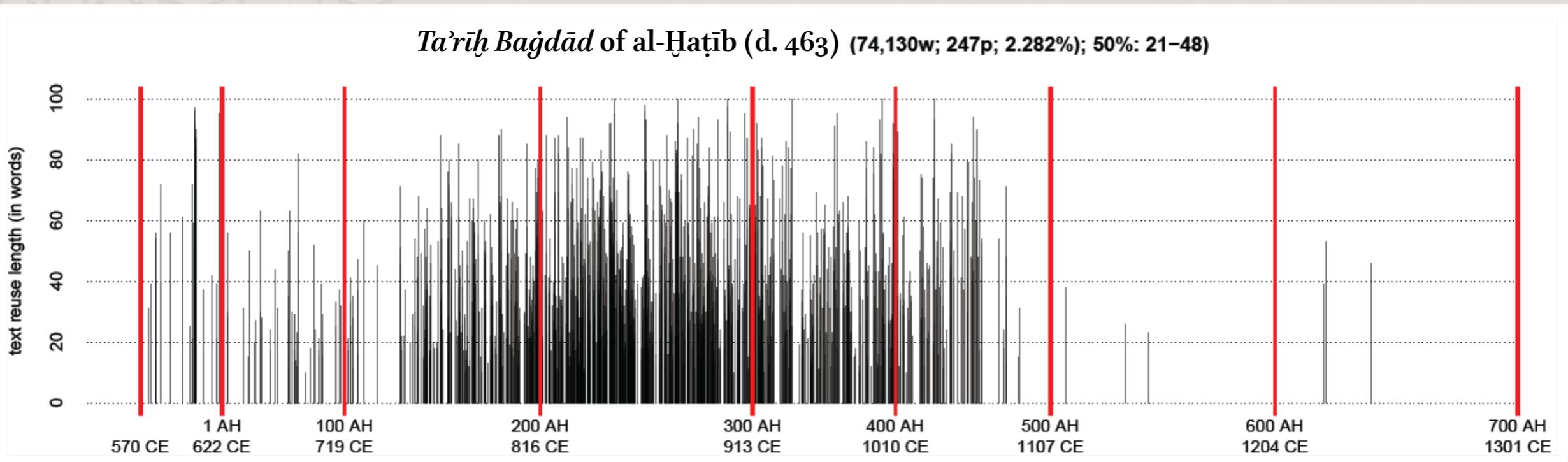


Text Reuse in the *Ta'rikh al-islam*

Dalā'il al-nubuwat of al-Bayhaqī (d. 458) (111,436w; 371p; 3.431%); 50%: 28–61)

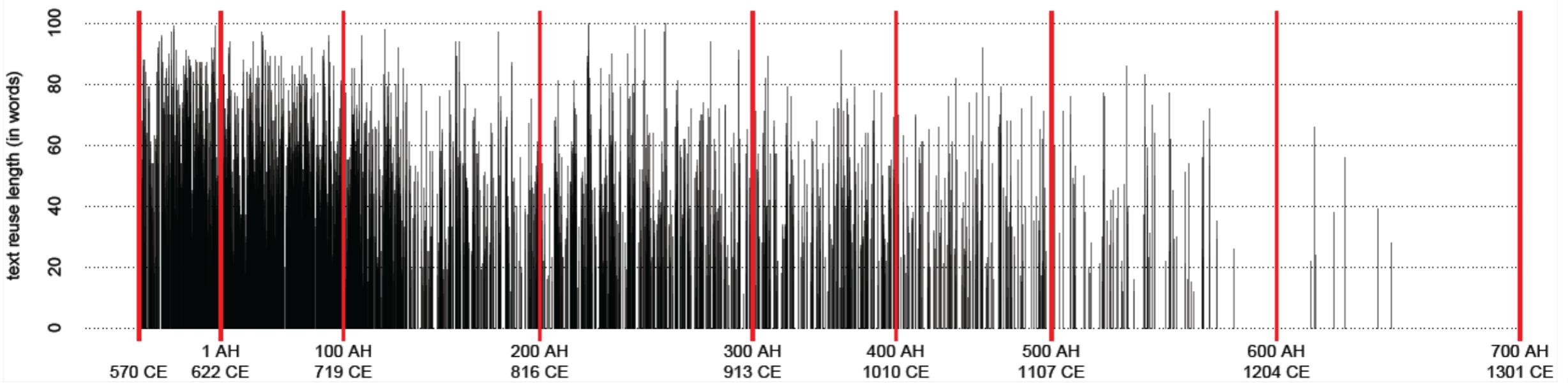


Ta'rikh Bağdād of al-Haṭīb (d. 463) (74,130w; 247p; 2.282%); 50%: 21–48)

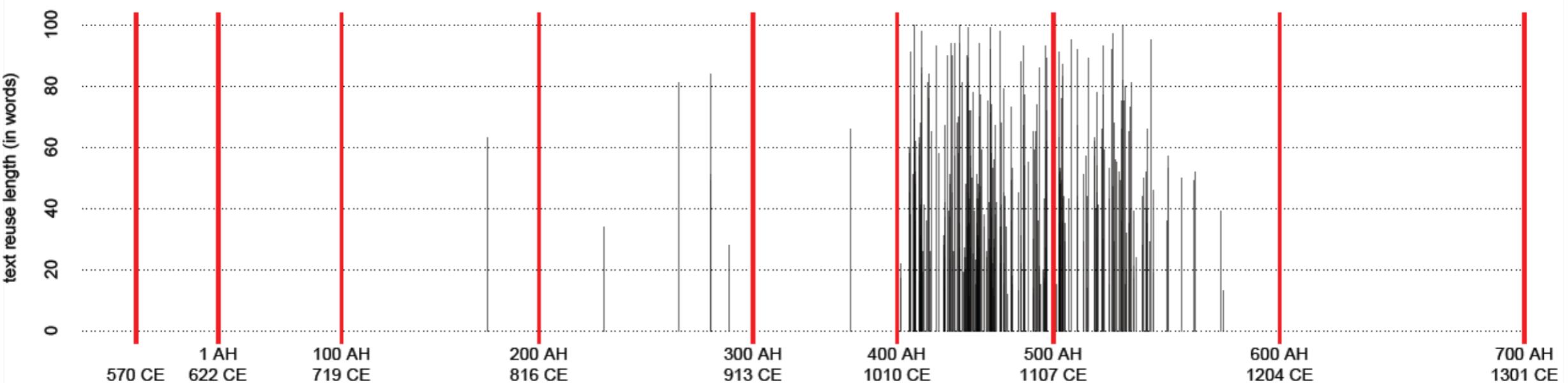


Text Reuse in the *Ta'rih al-islām*

Ta'rih Dimašq of Ibn 'Asākir (d. 571) [245,161w; 817p; 7.547%]; 50%: 22–48)

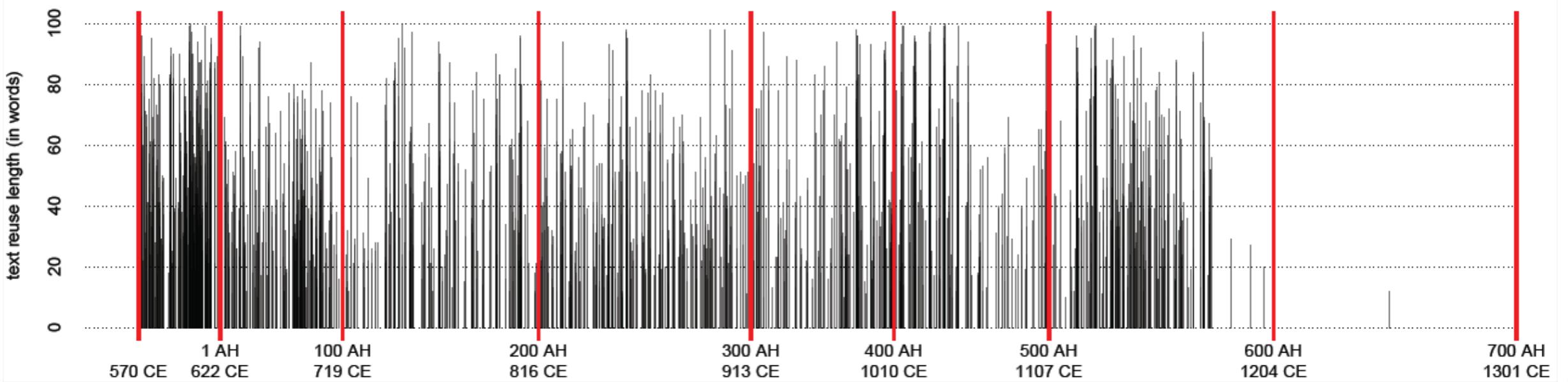


al-Silat of Ibn Baškuwāl (d. 578) (15,648w; 52p; 0.482%); 50%: 27–65)

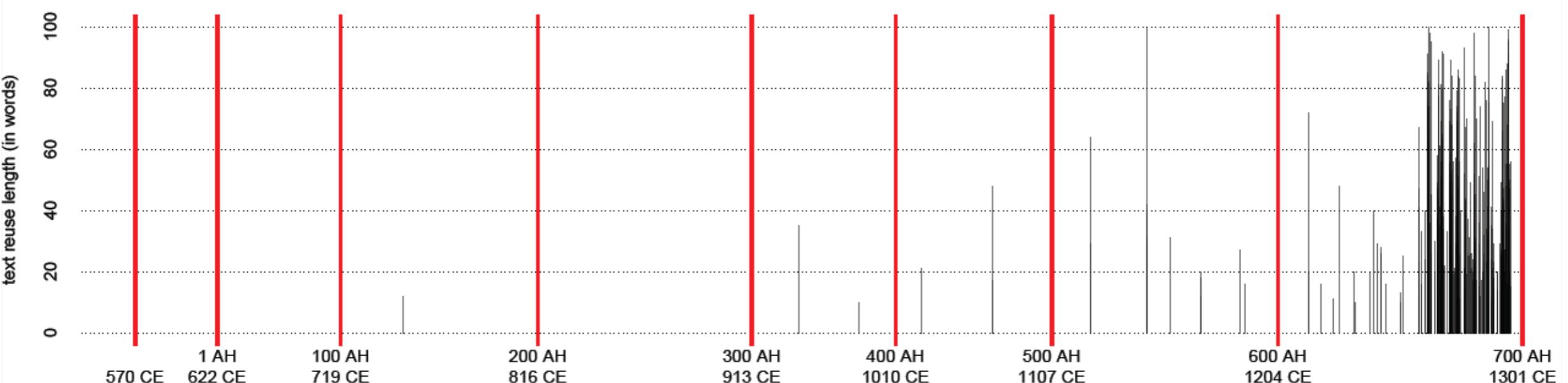


Text Reuse in the *Ta'rih al-islām*

al-Muntażam of Ibn al-Ǧawzī (d. 597) (83,828w; 279p; 2.581%); 50%: 25–60

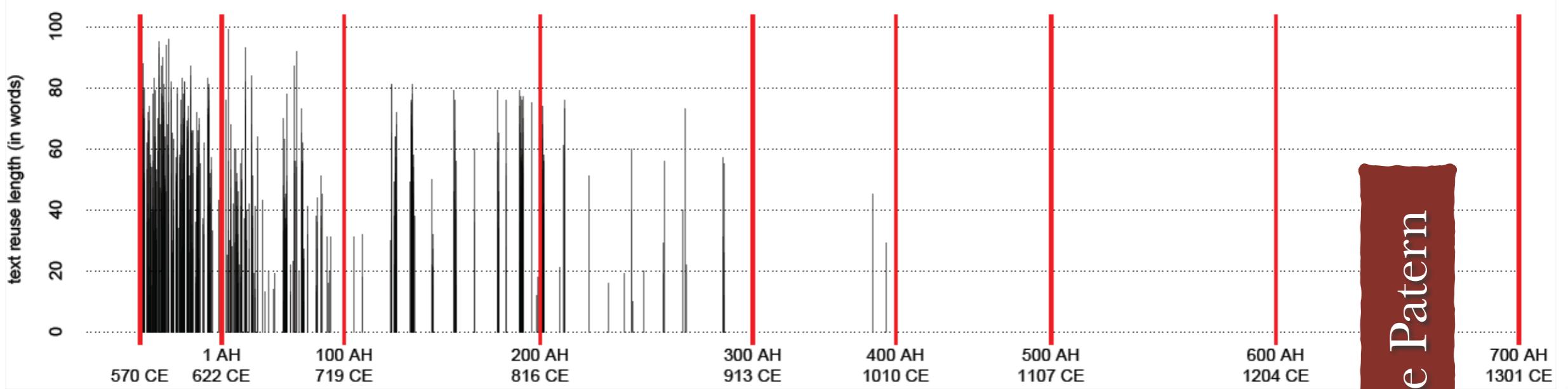


Dayl mir'at al-zamān of al-Yūnīnī (d. 726) (14,738w; 49p; 0.454%); 50%: 21–56



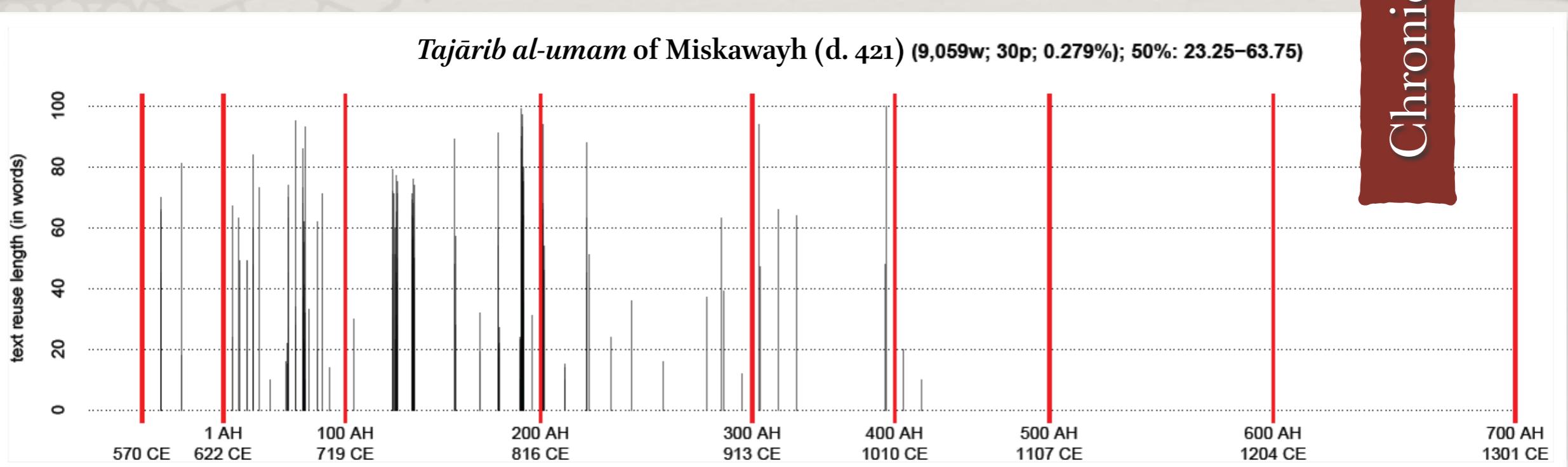
Text Reuse in the *Ta'rih al-islām*

al-Ta'rih of al-Ṭabārī (d. 310) (37,390w; 124p; 1.151%); 50%: 26–57)

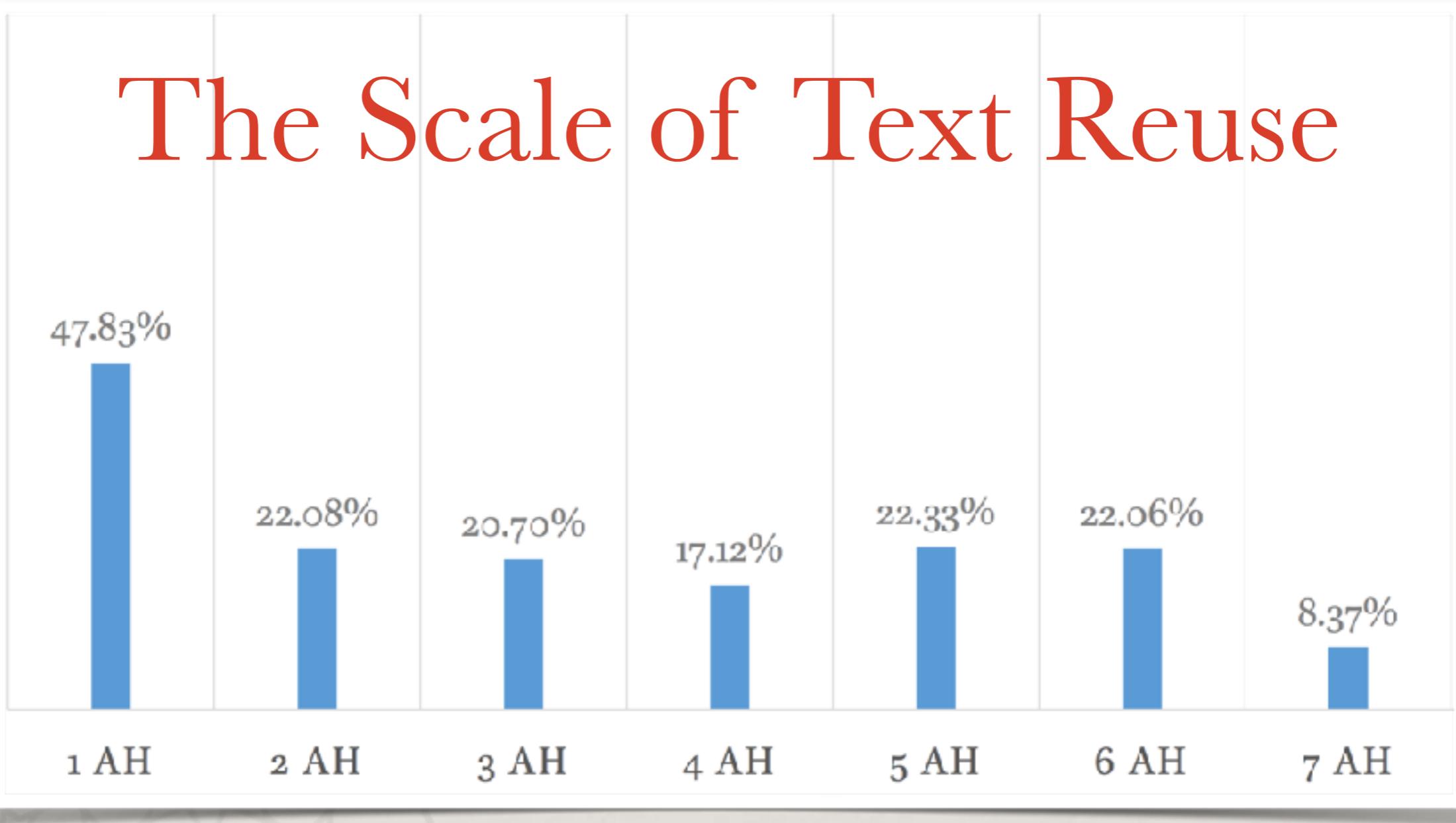


Chronicle Pattern

Tajārib al-umam of Miskawayh (d. 421) (9,059w; 30p; 0.279%); 50%: 23.25–63.75)



The Scale of Text Reuse



Counting only earliest quotations:

- 749,129 words
- 2,497 pages (300 w/p)
- $\sim 22.58\%$ (*unique*)
- 50% of quotes (25-59 words)

up to over 40.00%
(in current corpus)

The Scale: Distribution of Identifiable Borrowings (in pages [300 word units])

Coverage of different periods:	Sources traceable to different periods:							ALL
	1AH	2AH	3AH	4AH	5AH	6AH	7AH	
1 AH	0.75	68.56	537.74	86.27	52.43	35.05	6.06	786.87
2 AH	0.18	14.19	123.68	121.58	49.66	27.11	6.25	342.65
3 AH	0.08	2.96	95.56	93.91	113.49	34.70	9.57	350.26
4 AH	-	0.59	5.67	41.98	89.31	54.98	15.39	207.93
5 AH	0.11	1.20	5.07	13.27	75.04	112.22	54.11	261.03
6 AH	0.07	1.04	6.61	4.15	44.36	161.75	155.34	373.32
7 AH	-	0.51	5.04	6.45	20.83	6.94	135.25	175.04
ALL	1.20	89.06	779.37	367.62	445.13	432.75	381.97	2,497
								11,057

Note: This matrix shows that al-Dahabī is effectively using *archaic* language when he writes about the past: thus, when he writes about the 1st century AH (1st row), his narrative is dominated by quotations from text written in the 3rd century AH.



In progress: Social Network Analysis & Stylometric Analysis

Nodes: 7834 (99.72% visible)
Edges: 2558494 (100 % visible)

The Scale of Text Reuse

al-Dahabī's *Ta'rīh al-islām*



Social Network Analysis

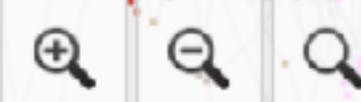
Nodes: 7834 (99.72% visible)
Edges: 2558494 (100 % visible)

The Scale of Text Reuse

Šī‘ī Tradition

al-Dahabī's *Ta'rīh al-islām*

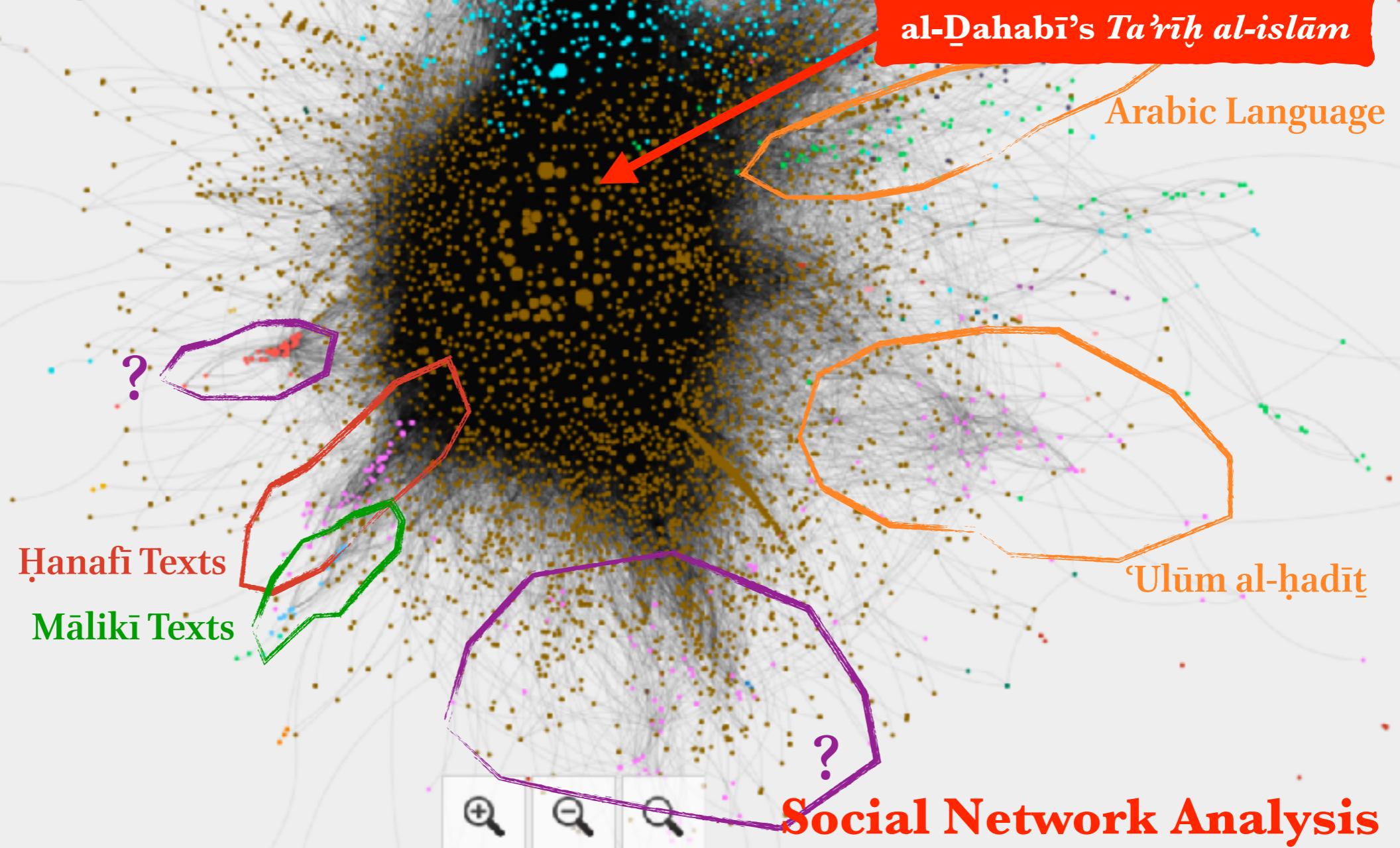
Sunnī Tradition



Social Network Analysis

Nodes: 7834 (99.72% visible)
Edges: 2558494 (100 % visible)

The Scale of Text Reuse



Stylometric Experiments

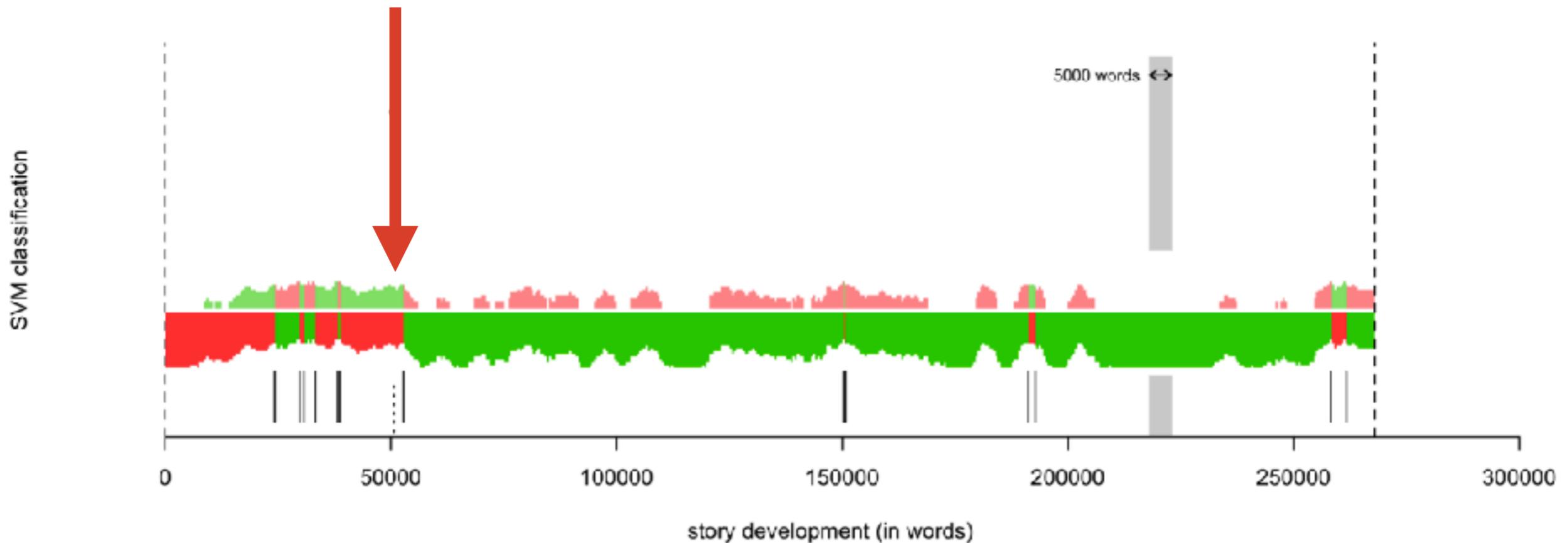
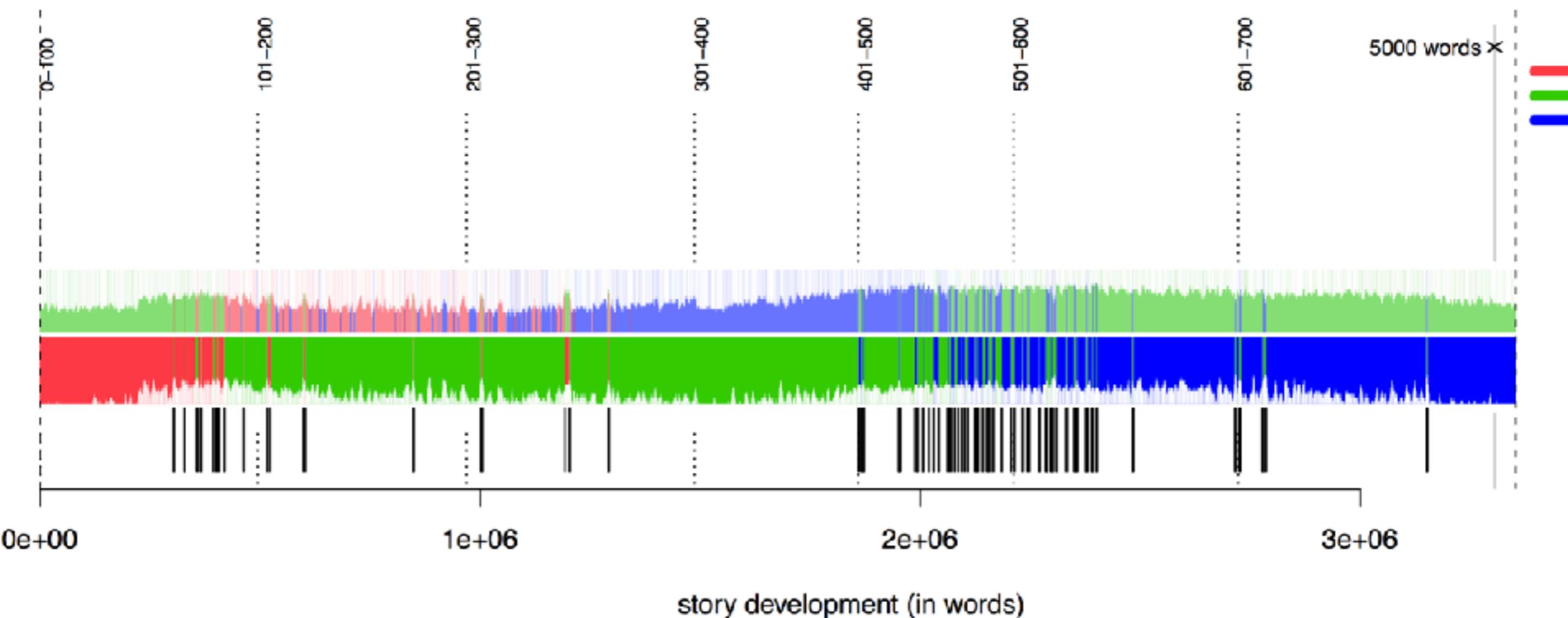


Fig. 1 Roman de la Rose assessed using Rolling SVM and 100 MFWs; window size: 5,000 words, sample overlap: 4,500 words. Sections attributed to Guillaume de Lorris are marked red, those attributed to Jean de Meun are green. The level of certainty of the classification is indicated by the thickness of the bottom stripe. The commonly-accepted division into two parts is marked with a vertical dashed line.

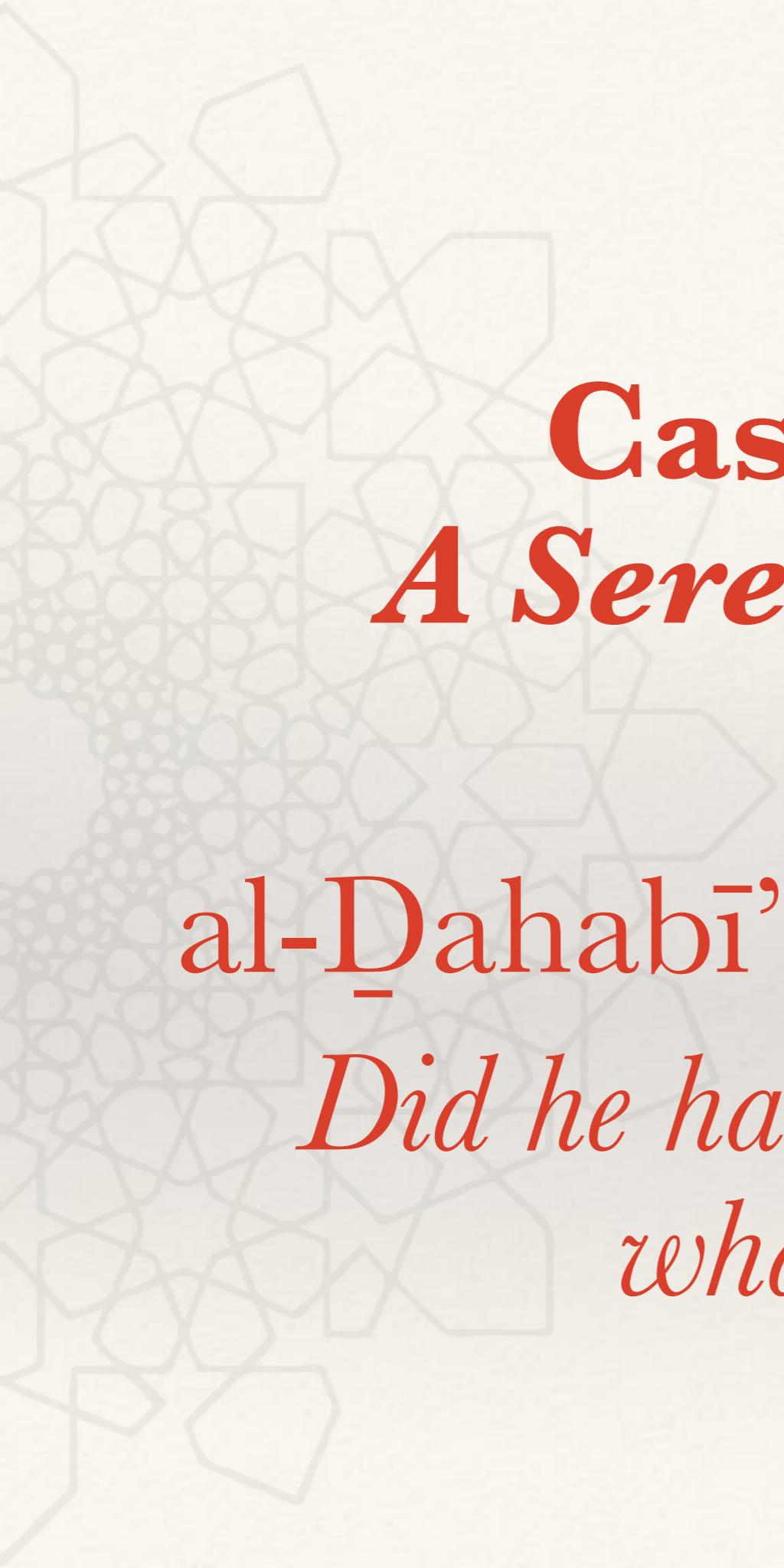
the 13th-century French allegorical poem *Roman de la Rose* attributed to **Guillaume de Lorris (red)** and **Jean de Meun (green)**

Stylometric Experiments



10,000 word sections from:
1st century AH, 4th century AH, 7th century AH

Testing al-Dahabī's “style” in the *Ta'rīh al-islām*



Case Study III: *A Serendipity Bonus*

al-Dahabī' Historical Method:
*Did he have any understanding
what he collected?*

al-Dahabī as a historian

- ❖ “Cities and Ports for Hearing the Reports”*
(al-Amṣār dawāt al-ātār)

* Translation by Michael Cooperson

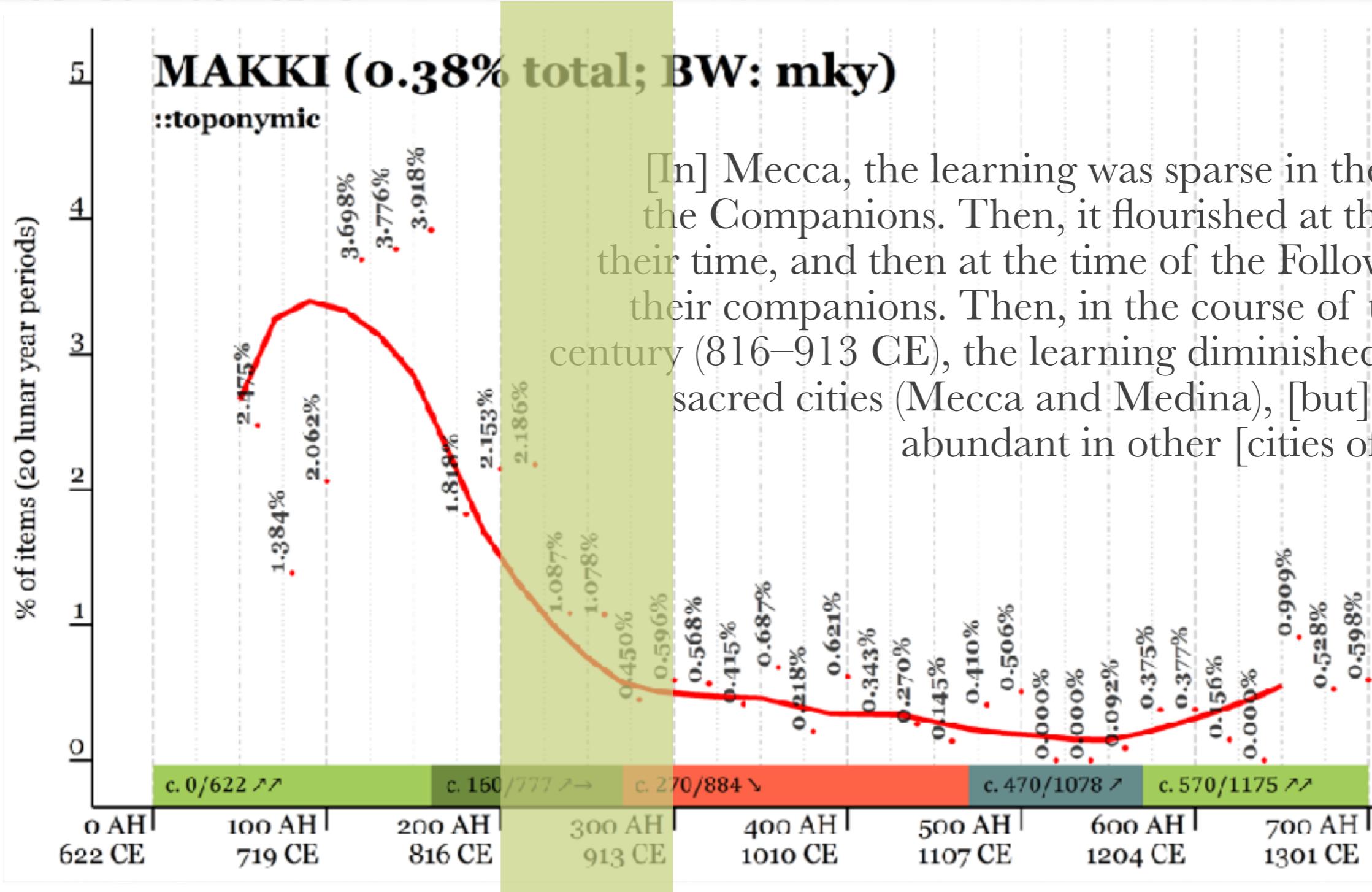
- ❖ 4-folio epistle
- ❖ focus on cities and provinces (~80)
- ❖ *statements about the role of the regions of the Islamic world in Hadīt learning*

Mecca / Makkat

مكة ... كان العلم بها يسيرا في زمن الصحابة
ثم كثُر في أواخر عصر الصحابة
وكذلك في أيام التابعين وزمن أصحابهم ...
ثم في أثناء المائة الثالثة تناقص علم الحرمين وكثُر بغيرهما

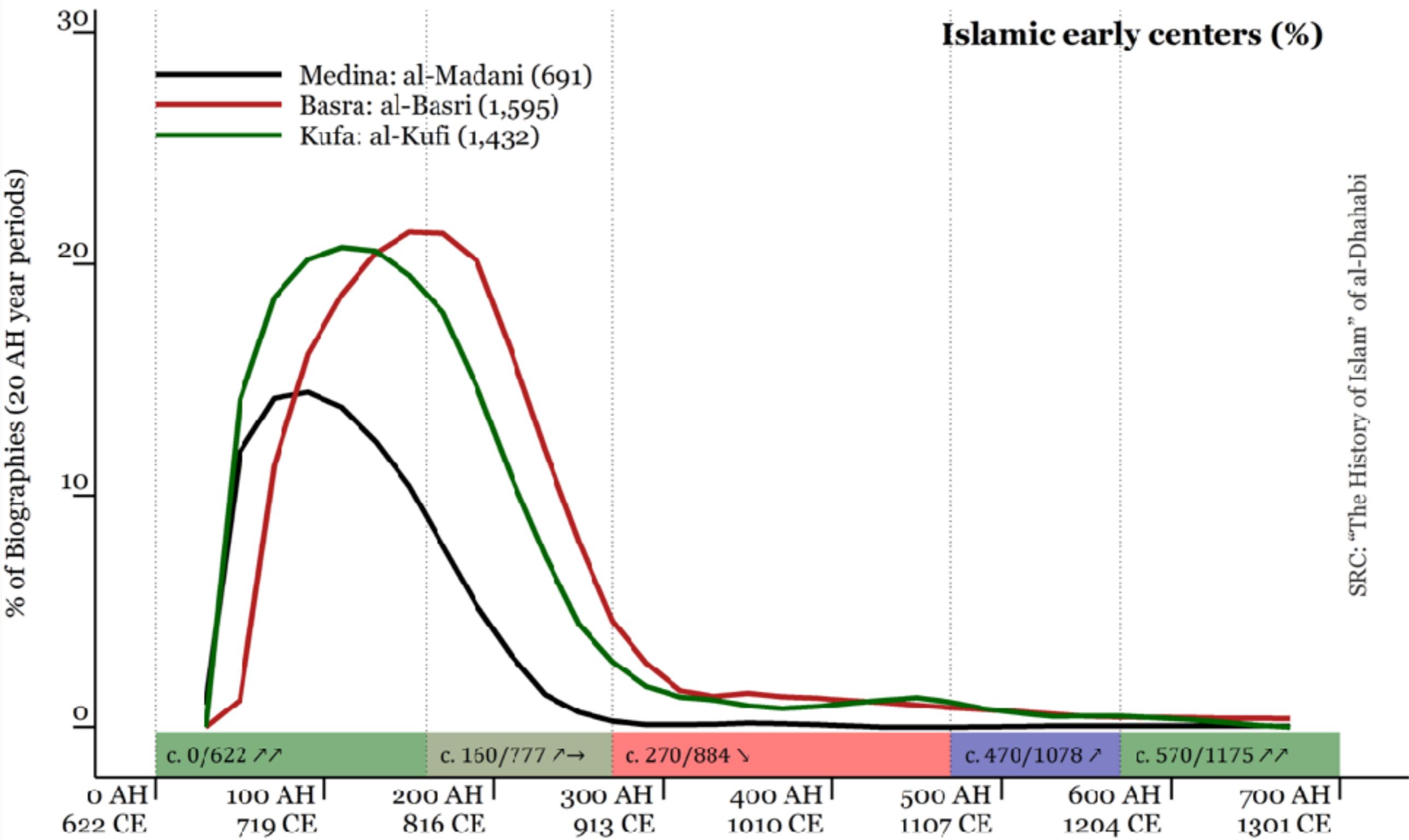
[In] Mecca, the learning was sparse in the time of the Companions. Then, it became abundant at the end of their time, and then in the time of the Followers and their companions. Then, in the course of the third century (816–913 CE), the learning diminished in the two sacred cities (Mecca and Medina), [but] became abundant in other [cities of Islam].

Mecca / Makkat

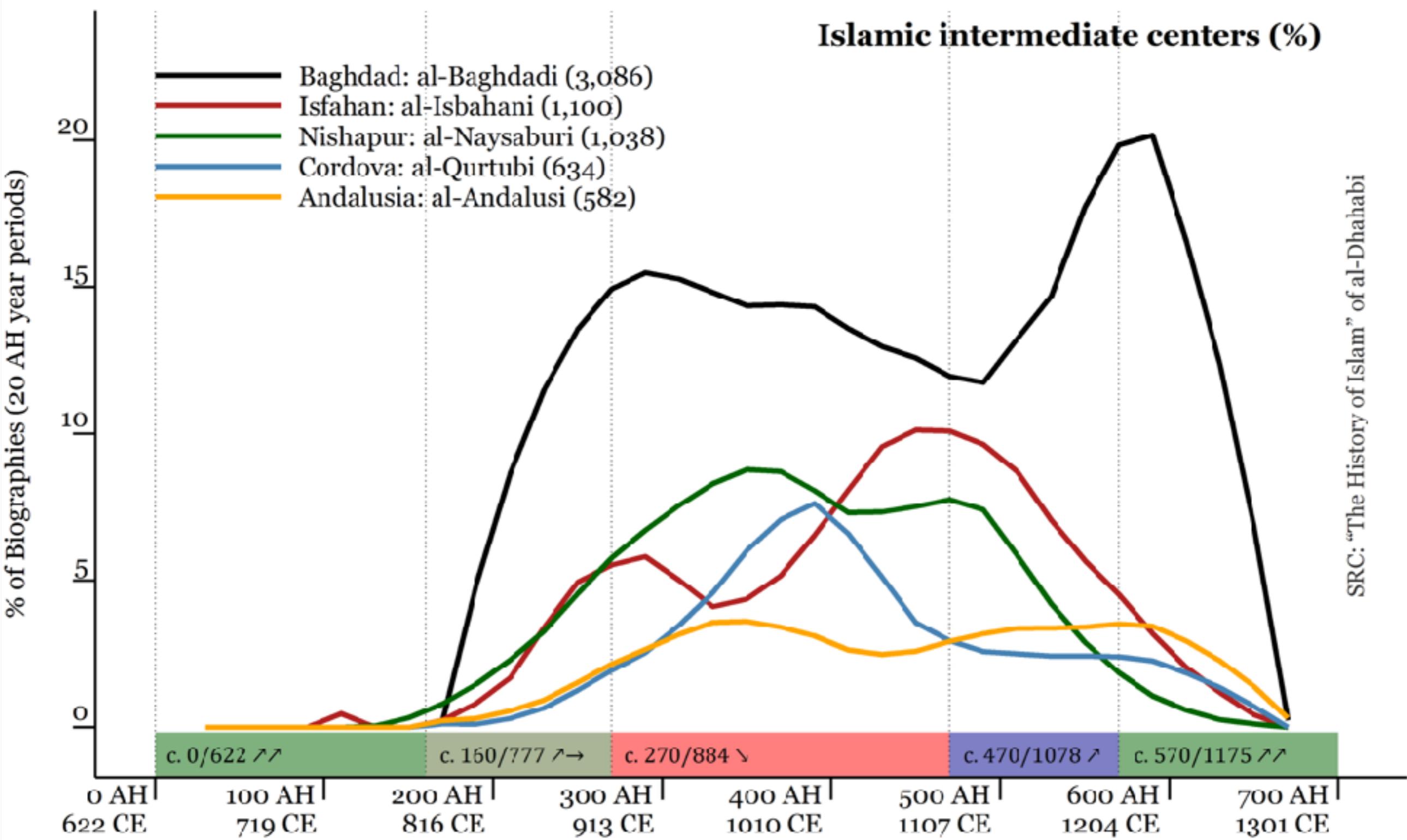


[In] Mecca, the learning was sparse in the time of the Companions. Then, it flourished at the end of their time, and then at the time of the Followers and their companions. Then, in the course of the third century (816–913 CE), the learning diminished in both sacred cities (Mecca and Medina), [but] became abundant in other [cities of Islam].

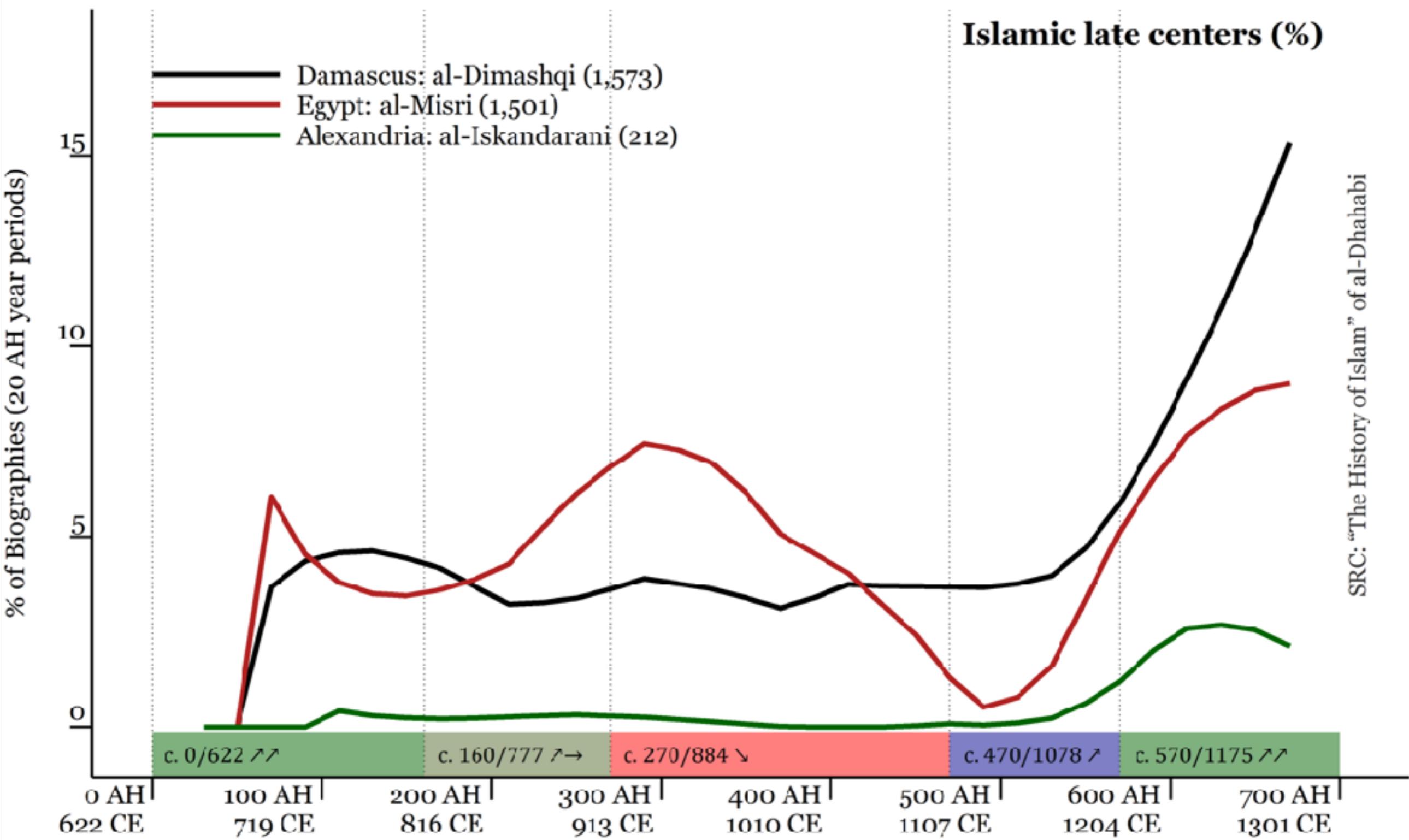
Islamic early centers (%)



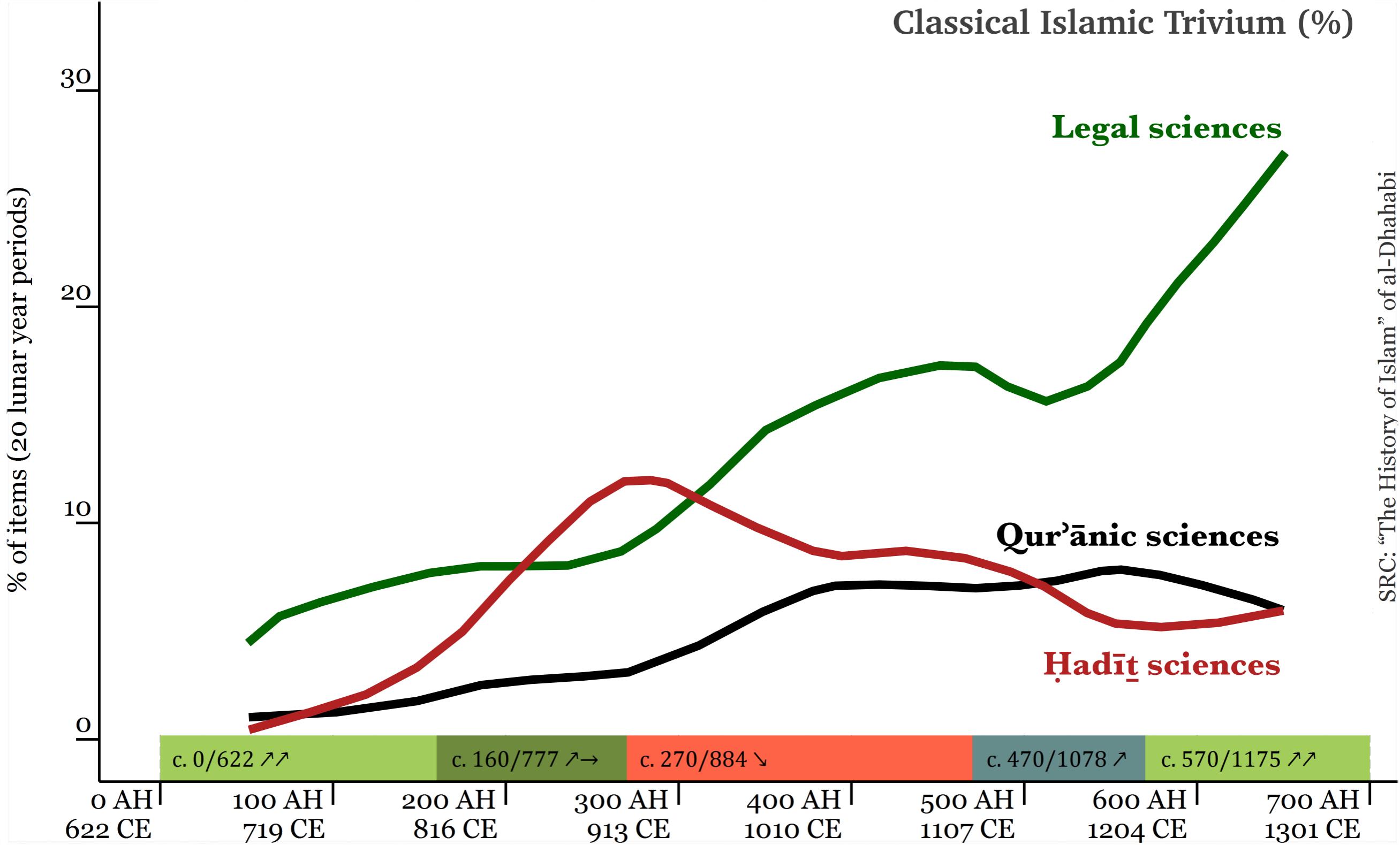
Islamic intermediate centers (%)



Islamic late centers (%)



Classical Islamic Trivium (%)



al-Amsār dawāt al-ātār

- ❖ **Certain statements**
about ***large groups*** in the *Ta'rih al-islām*
(at least 100 people, but usually much more)
- ❖ **Uncertain statements**
about ***small groups*** in the *Ta'rih al-islām*
- ❖ **No statements**
no identifiable groups in the *Ta'rih al-islām*

What was his method? Did he count?!



al-Dahabī's method: How?



Collecting,
Organizing,
Categorizing

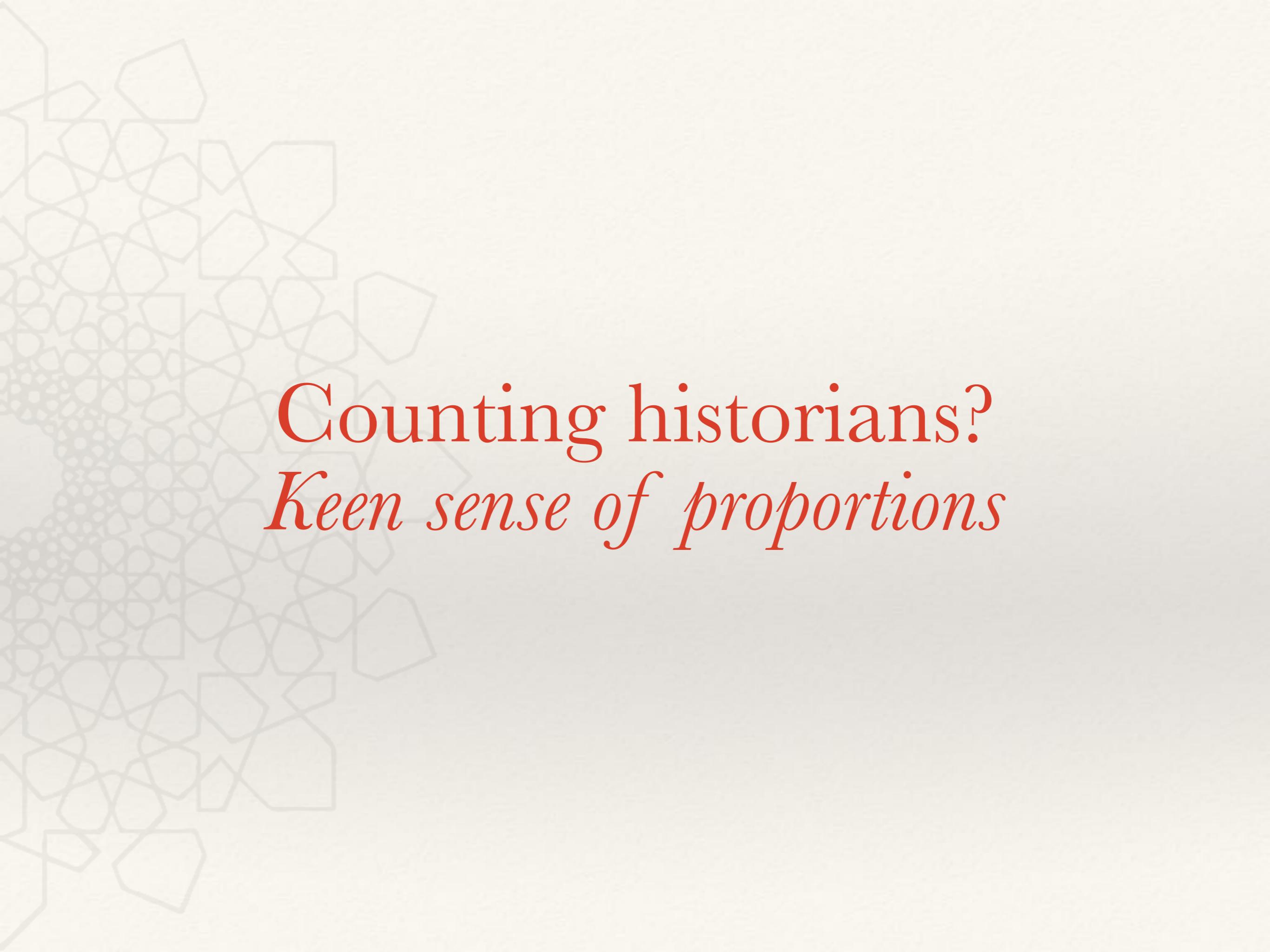
Collecting,
Organizing,
Categorizing,
Re-organising

Counting Muslims

- Hadith collections
- Lexicographical dictionaries
- Onomastic dictionaries
- Terminological dictionaries
- Genealogical texts
- Biographical collections
- Geographical texts
- Chronicles

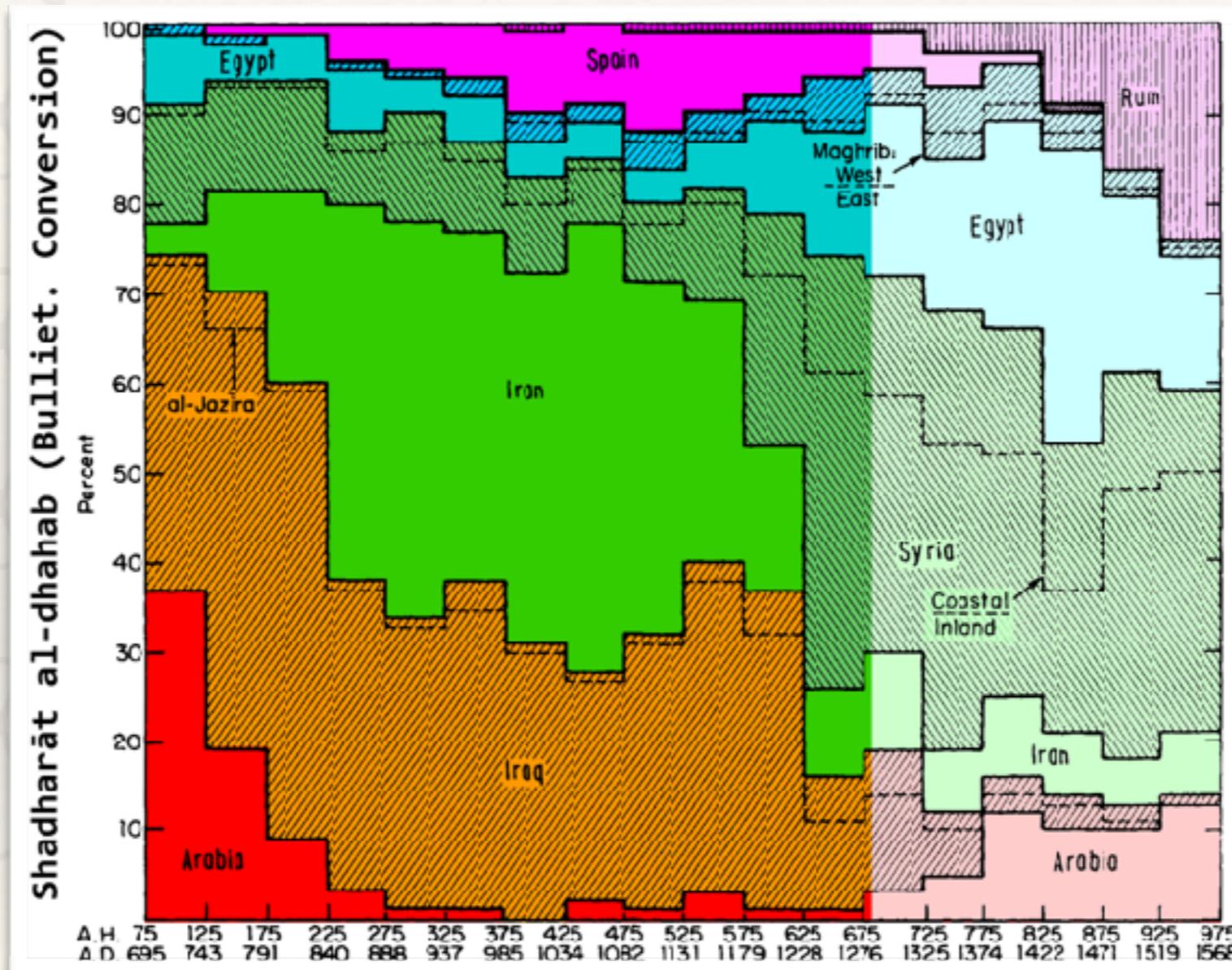
*Texts with serialized data
run into hundreds*

Plus: evidence of awareness of statistical thinking!

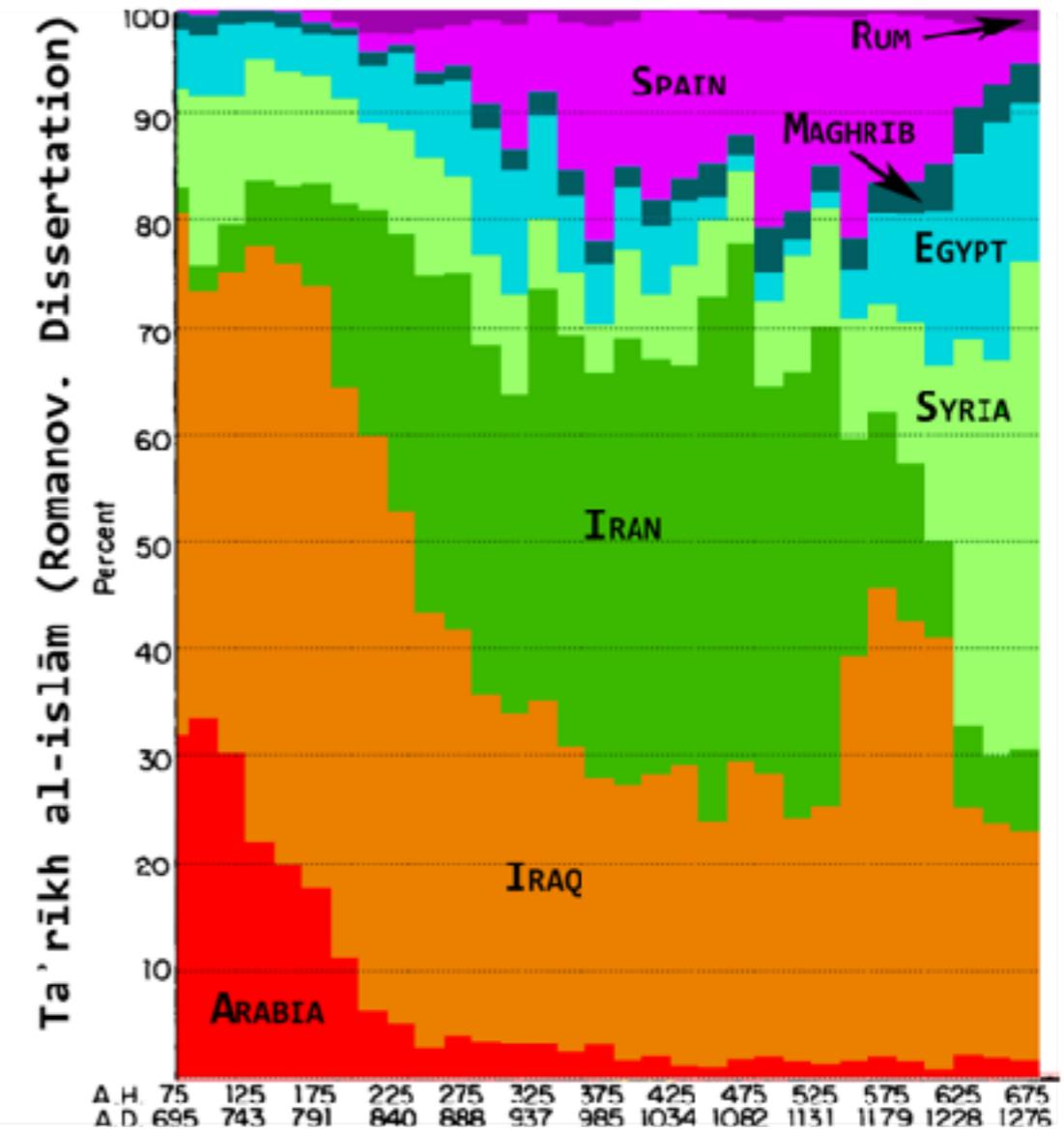
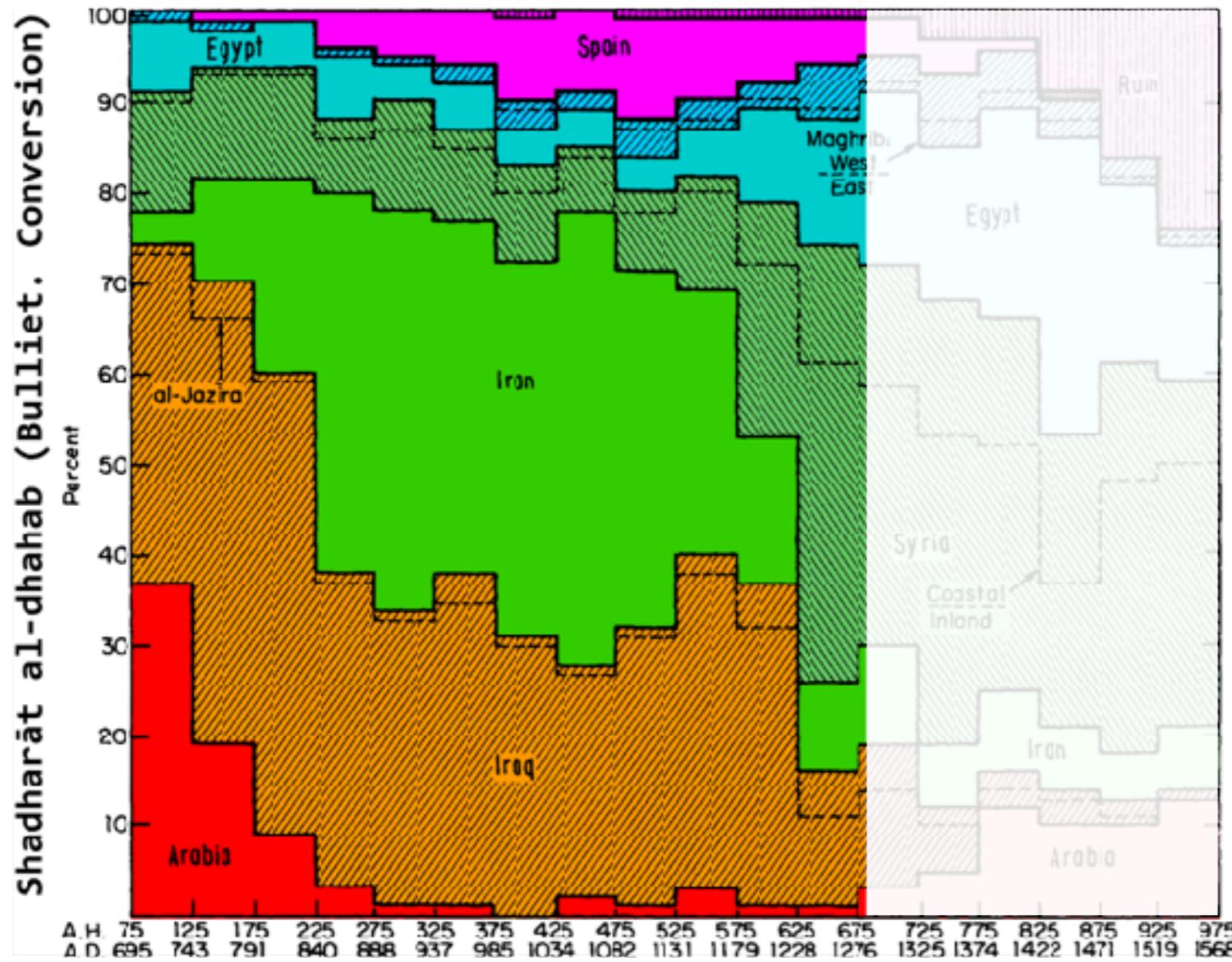


Counting historians?
Keen sense of proportions

Geo-Chronological Coverage: Comparative Perspective

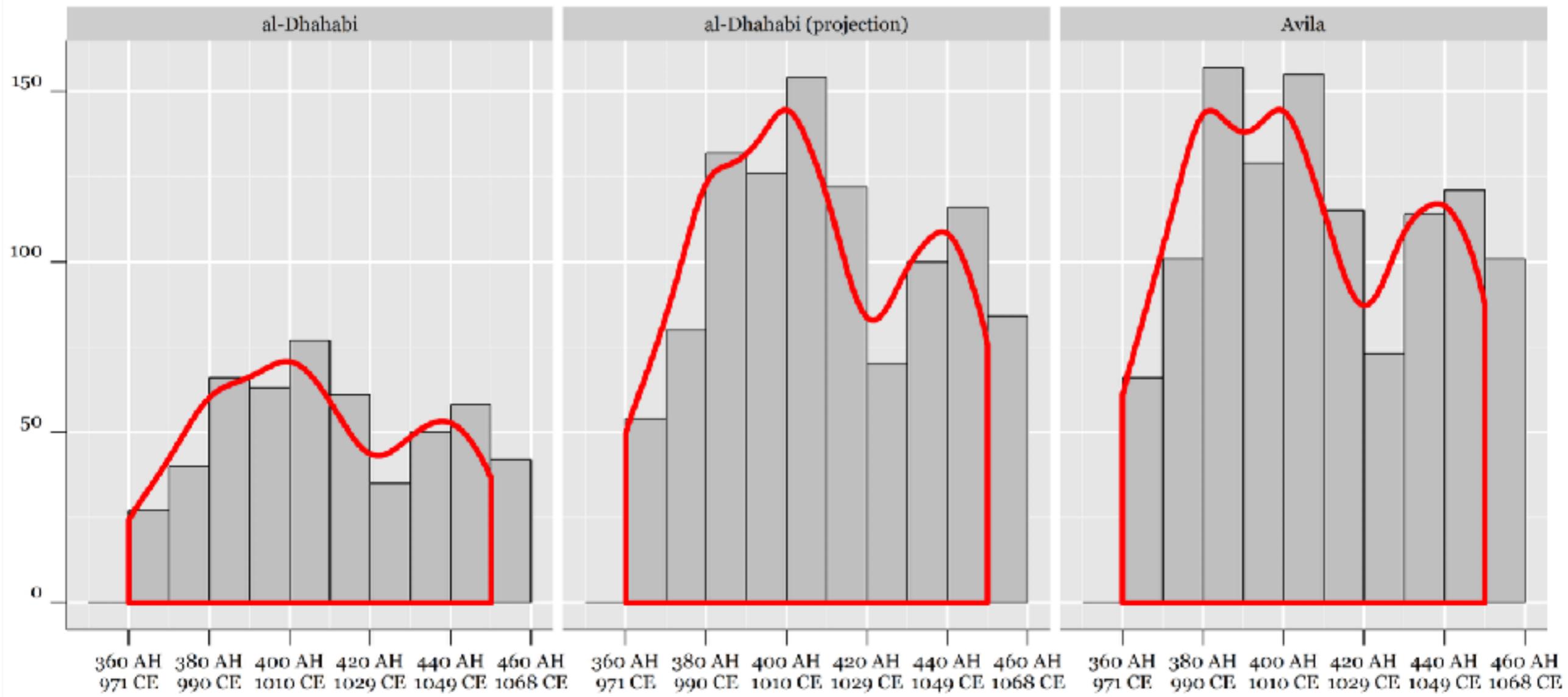


Geo-Chronological Coverage: Comparative Perspective



Geo-Chronological Coverage: Comparative Perspective

AH	370	380	390	400	410	420	430	440	450	460
in TI	42%	40%	42%	49%	49%	53%	49%	44%	48%	41%





Working with data?

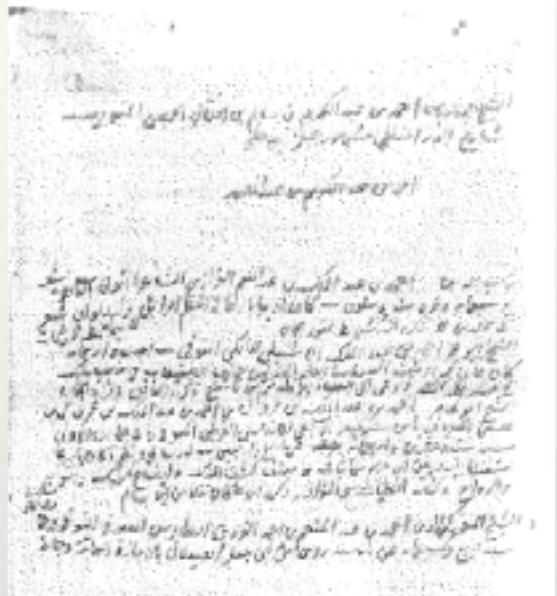
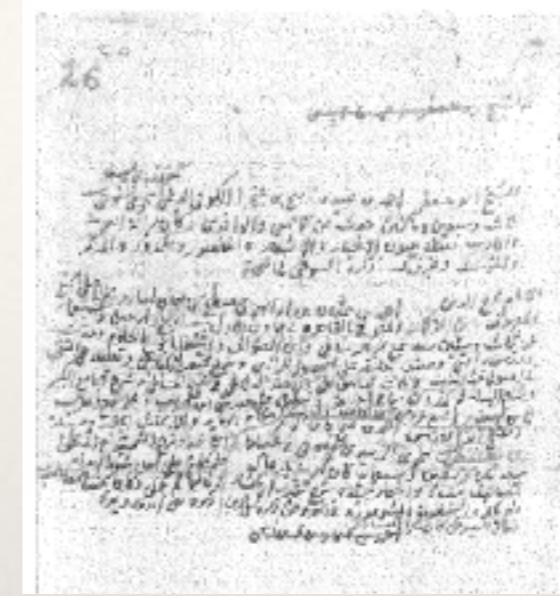
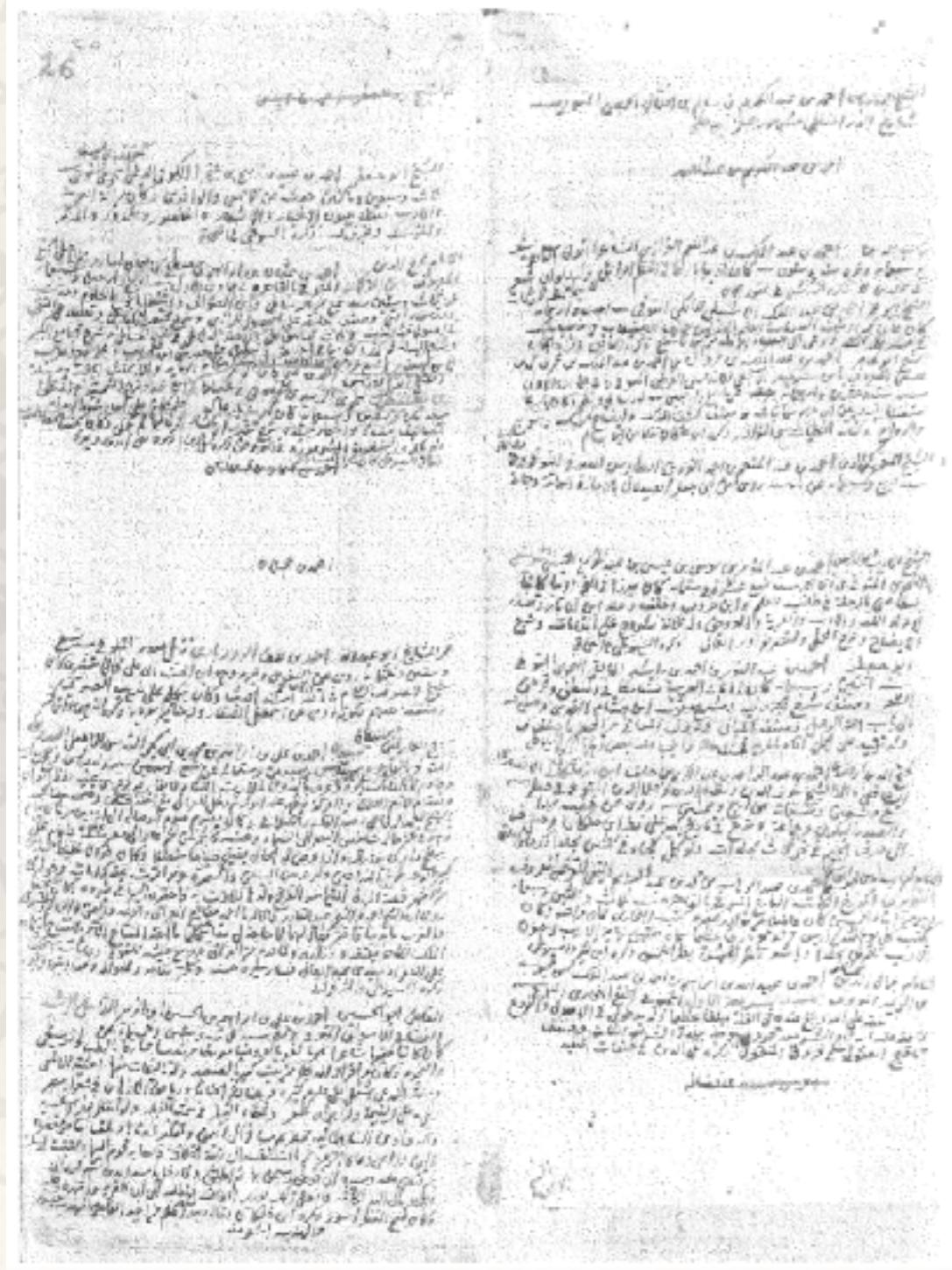
Movable media



Reference:

Birnbaum, E. “Kātib Chelebi (1609-1657) and Alphabetization: A Methodological Investigation of the Autographs of His *Kashf al-zunūn* and *Sullam al-wusūl*.” In *Scribes et Manuscrits Du Moyen-Orient*. Sous La Dir. de F.Déroche & F.Richard, 235–63. Bibliothèque Nationale de France, 1997.

Movable media



Movable media



Movable media

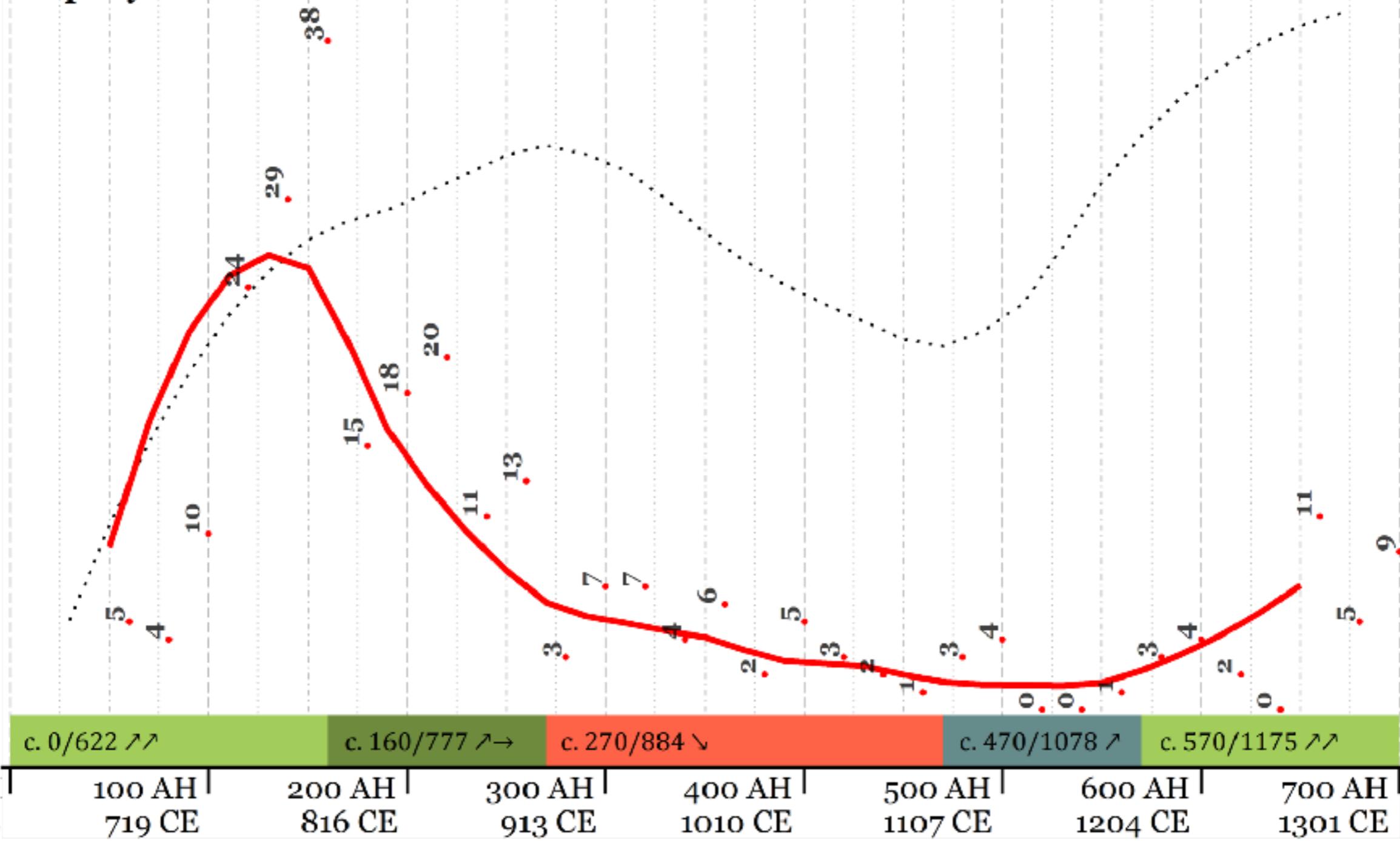


Movable media

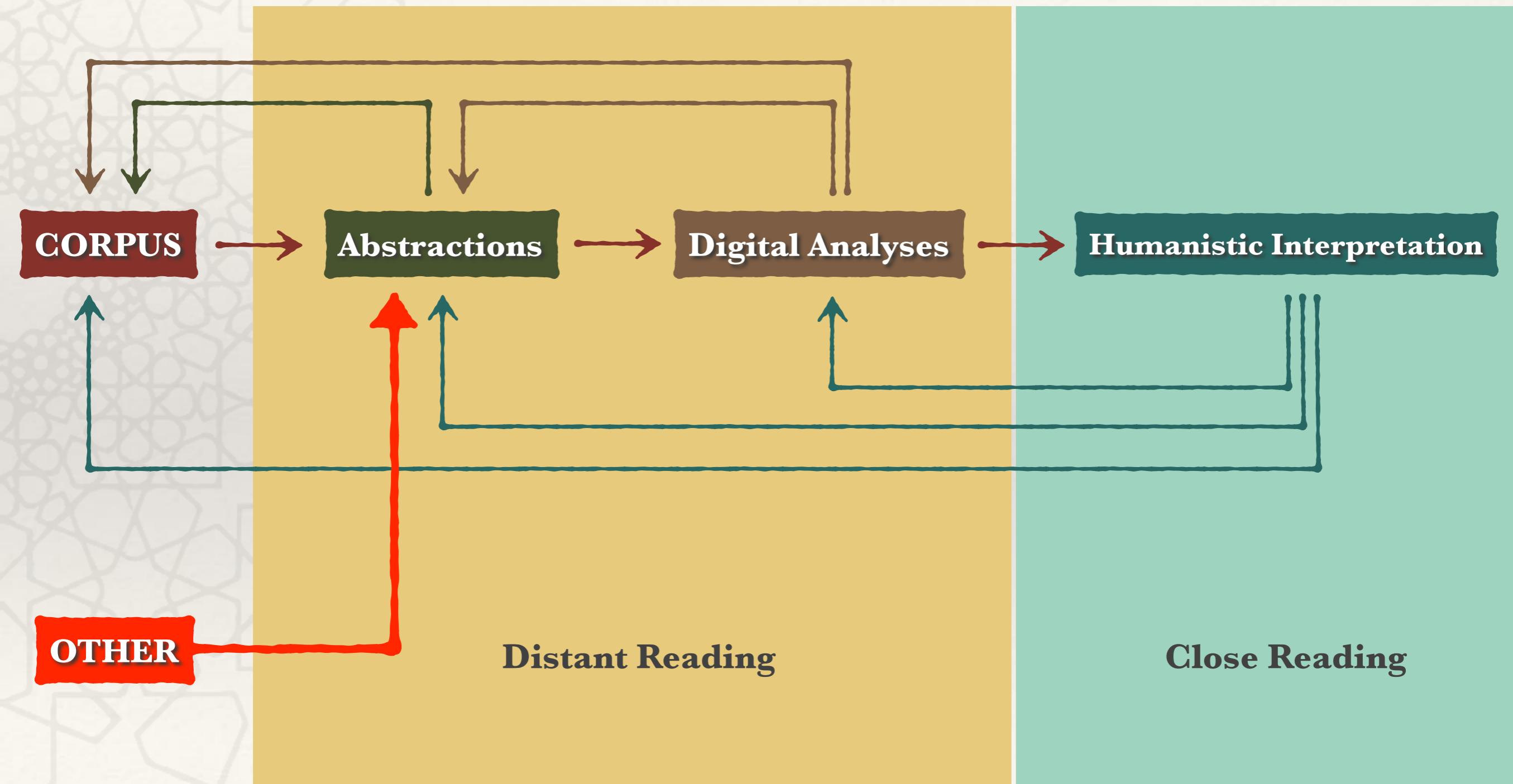
MAKKI (269 total; BW: mky)

..... Cumulative biographical curve (29,110; scaled down by a factor of 38.03)

::toponymic



Conclusion



There's a graph
for that

