Catch me if you can! Tracing the late Ottoman ideosphere through network analysis and stylometry of the Arabic periodical press

Till Grallert, Orient-Institut Beirut (OIB), @tillgrallert

3rd Online Conference of the Islamicate Digital Humanities Network (IDHN)

29 April 2020

Slides: https://OpenArabicPE.github.io/slides/2020-idh/

Overview

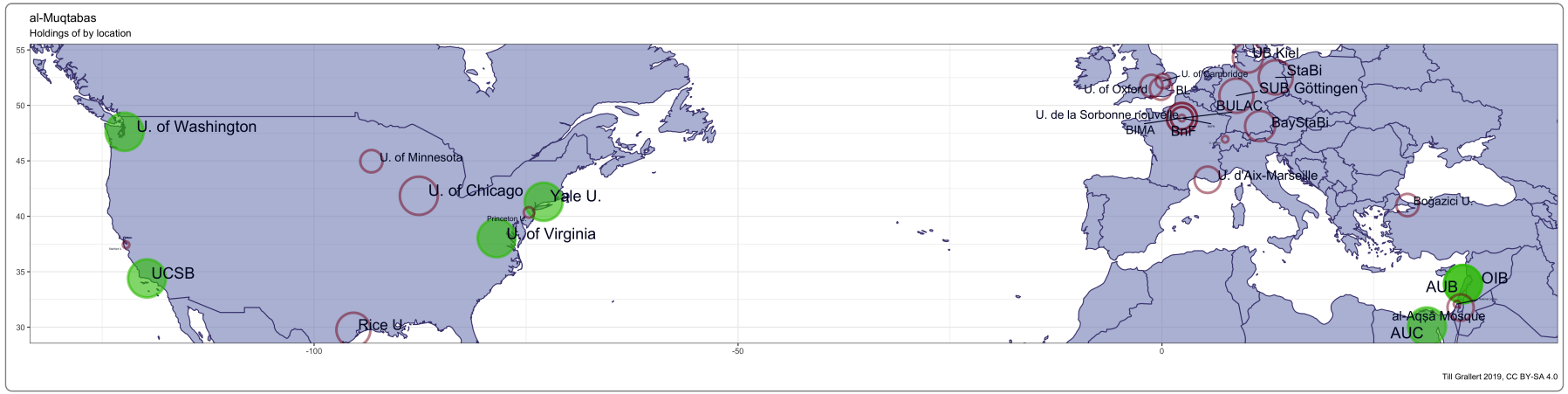
- Problematic state of Arabic periodical studies
- Methodologic proposals to computationally address this state
- My approaches
    - bibliometrics
    - (social) network analysis
    - stylometric authorship attribution

Introduction

Problems

1. periodicals are commonly perceived as a *source* but not a *subject* in its own right
2. the history of the periodical press
    - focusses on a few “core” publications from Cairo and Beirut
    - remains largely unexplored
    - is full of untested hypotheses
    - is heavily biased by national(ist) narratives

One of the many reasons



Map: geographic distribution of library holdings of *al-Muqtabas*

Open questions:

micro level

- how were individual periodicals produced?
- who authored the vast majority of anonymous articles?

meso level

- What are the core nodes (authors, periodicals, other works) in the ideosphere of the late Ottoman Eastern Mediterranean?

macro level

- How do we have to revise the *nahḍa* narrative / late Ottoman intellectual history if we include the full ideosphere of the press beyond Cairo and Beirut?

Methodology

Methodology: computational approaches

- *distant reading*
    - bibliometrics
    - social network analysis
    - stylometric authorship attribution
- requires a digital corpus -> *corpus building*

corpus building

It's *labour and resource intensive*. It really is!

1. get the data
    - text: transcription, train OCR/HTR
    - facsimiles: scanning
    - bibliographic metadata: transcription, validated iterative generation
2. transform the data into a human and machine readable edition
    - model the source
    - identify entities and link them to authority files
3. host, share and preserve the data

corpus building: Open Arabic Periodical Editions

1. ideas:
    - unite *available* facsimiles and transcriptions in a standard-compliant format
    - harvest, generate, validate and share open metadata
2. aims
    - *validate* the transcription against the facsimiles
    - *improve* the transcription with the help of the “crowd”
    - make everything *citable* for scholars, *linkable* for machines
    - share all data, metadata and tools with the broadest possible licences to facilitate access and re-use
3. principles
    - re-purpose *available* and *established* tools, technologies, and material
    - preference for *open* and *simple* formats and tools

corpus building: Open Arabic Periodical Editions

1. Open licences: CC BY-SA 4.0 (TEI, MODS, BibTeX), MIT license (XSLT, XQuery)
2. Social digital editions hosted on GitHub: gradually improve transcription and mark-up
3. Releases are archived at Zenodo: receive a DOI for reliable citation
4. Static web-view: provides side-by-side view of facsimiles and text
5. Access to bibliographic metadata through a public Zotero group

OpenArabicPE's corpus

periodical	doi	volumes	issues	articles	words	words per article
al-Ḥaqā'iq	10.5281/zenodo.1232016	3	35	389	298090	832.66
al-Ḥasnā'	10.5281/zenodo.3556246	1	12	201	NA	NA
al-Manār		35	537	4300	6144593	1437.73
al-Muqtabas	10.5281/zenodo.597319	9	96	2964	1981081	873.34
al-Ustādh	10.5281/zenodo.3581028	1	42	435	221447	582.21
al-Zuhūr	10.5281/zenodo.3580606	4	39	436	292333	695.09
Lughat al-ʿArab	10.5281/zenodo.3514384	3	34	939	373832	485.21
total		56	795	9664	9311376	

OpenArabicPE's corpus

title	articles with author (%)	authors	author.birth.avg	author.death.avg
al-Ḥaqā'iq	41.90	104	1837.53	1905.93
al-Ḥasnā'	37.31	50	1883.14	1947.17
al-Manār	*87.14*	356	1870.96	1913.5
al-Muqtabas	12.72	140	1855.94	1929.48
al-Ustādh	5.52	8	NA	NA
al-Zuhūr	41.51	112	1872.96	1939.52
Lughat al-ʿArab	16.19	53	1875.8	1942.1

First attempts to computationally map the ideosphere of the late Ottoman Arabic-speaking press

Evaluating the corpus: network of referenced periodicals

network of referenced periodicals: data sources

- mark-up of all references to periodicals in the text
  - semi-automatic
- authority files for disambiguation and additional information
  - mostly automatic

```
(راجع والأصح الدرعية بلام التعريف <bibl subtype="journal"
        type="periodical">مجلة <title level="j"
        ref="oape:bibl:3 oclc:1034545644"
        xml:id="title_16.d2e2291">الزهور</title> المصرية
        <biblScope unit="volume" from="2"
        to="2">٢</biblScope> : <biblScope unit="page"
        from="292">٢٩٢</biblScope></bibl>)
```

```
وانتخب <persName>فؤاد أفندي الدفتري البغدادي</persName>
        و<bibl><editor><persName>نوري أفندي</persName>
        </editor> راس كتاب <textLang otherLangs="ota">القسم
        التركي</textLang> في <bibl type="periodical"
        subtype="newspaper">جريدة <title
        ref="oape:bibl:532">الزهور</title></bibl>
        البغدادية</bibl> نائبين عن <placeName
        ref="oape:place:372 geon:94824">كربلاء</placeName>.
```

network of referenced periodicals

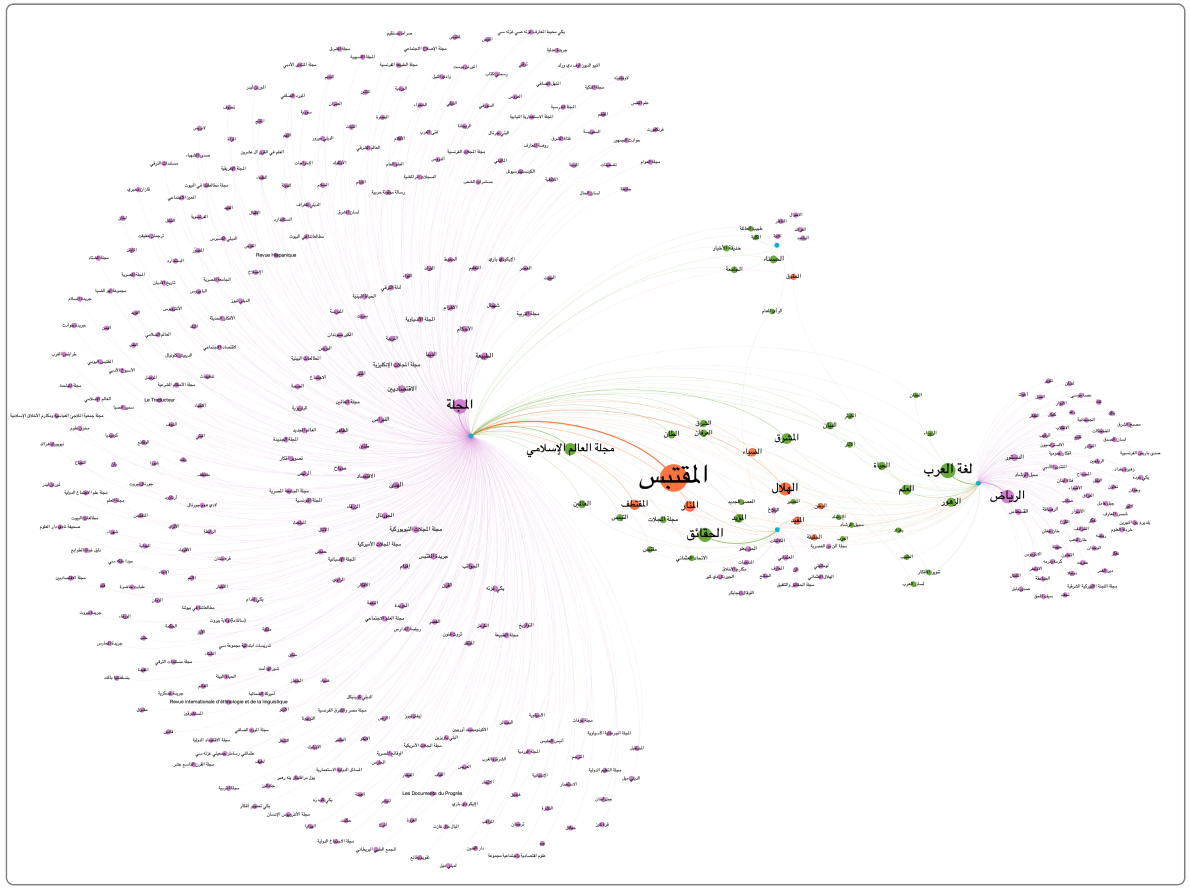


Figure: Network of periodicals mentioned *al-Ḥaqā'iq*, *al-Ḥasnā'*, *Lughat al-'Arab* and *al-Muqtabas*; weights per issue

1. only a few nodes are of relative importance (44 of 465)
2. *al-Muqtabas* accounts for the vast majority of references
3. all periodicals are primarly self-referential

network of referenced periodicals: the core

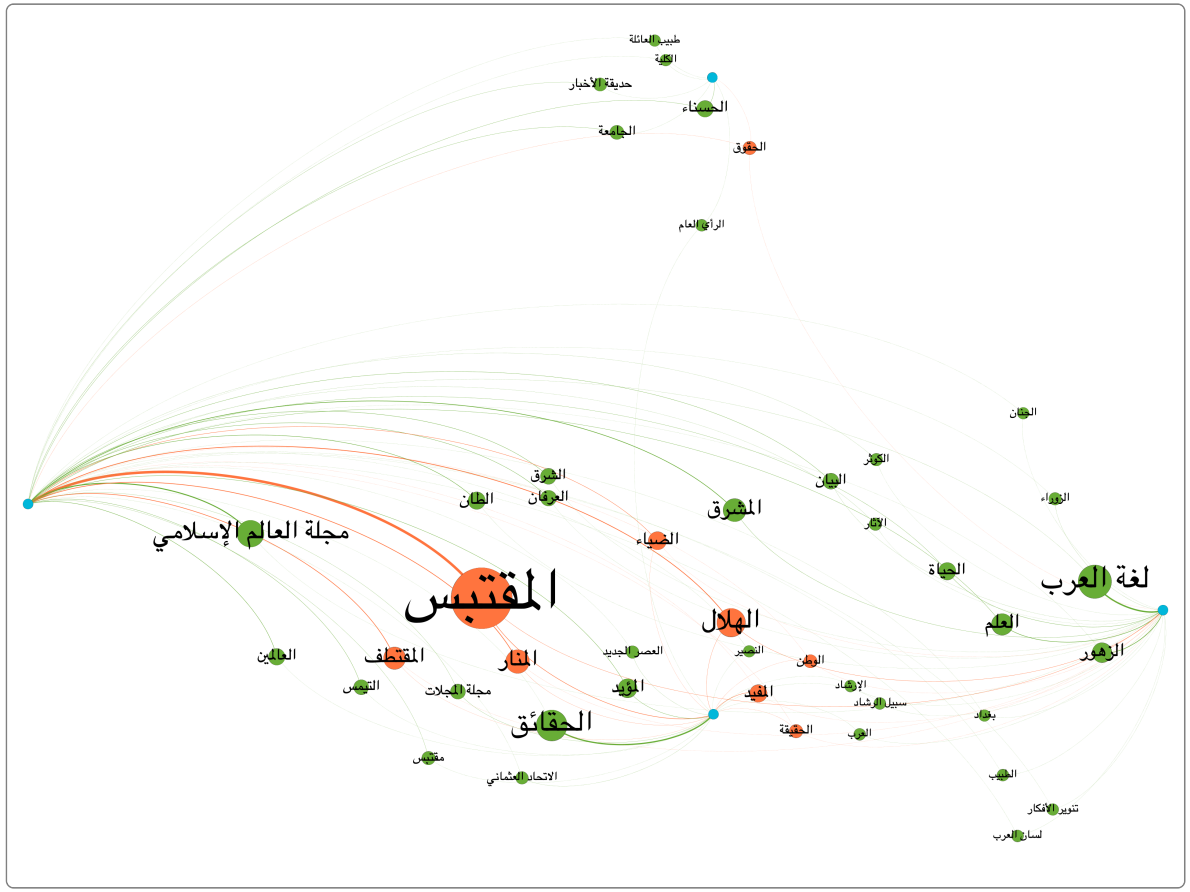


Figure: Core nodes in the network of periodicals mentioned *al-Ḥaqā'iq*, *al-Ḥasnā'*, *Lughat al-'Arab* and *al-Muqtabas*; weights per issue

1. confirms the bias on Cairo and Beirut (8 of 9)
2. highly centralised in terms of geographic distribution (10 locations)
3. surprising members

Bibliometrics: the network of authors

network of authors: data sources

- structured bibliographic data provided by the text itself
  - semi-automatic
  - problems: many accronyms, plurality of name forms
- authority files for disambiguation and additional information
- automatic enriching: from semantic web
  - life dates
  - works
  - geocoded locations

```
<person>
    <persName><roleName type="pseudonym">ساتسنا</roleName>
        </persName>
    <persName><roleName type="pseudonym">أمكح</roleName>
        </persName>
    <persName><roleName type="pseudonym">فهر
        الجابري</roleName></persName>
    <persName><roleName type="rank">الأب</roleName>
        <forename>أنستاس</forename>
        <forename>ماري</forename> <surname><addName
        type="nisbah">الكرملي</addName></surname></persName>
    <persName><forename>أنستاس</forename>
        <forename>ماري</forename> <addName
        type="nisbah">الألياوي</addName> <surname><addName
        type="nisbah">الكرملي</addName></surname></persName>
    <persName><forename>بطرس</forename> <addName
        type="nasab">بن <forename>جبرائيل</forename>
        </addName> <forename>يوسف</forename>
        <surname>عواد</surname></persName>
    <idno type="VIAF">39370998</idno>
    <idno type="oape">227</idno>
    <idno type="wiki">Q4751824</idno>
    <birth><date source="viaf" when="1866-08-05">1866-08-
        05</date> in <placeName ref="oape:place:216
        geon:98182">Baghdad</placeName></birth>
    <death><date source="viaf" when="1947-01-07">1947-01-
        07</date> in <placeName ref="oape:place:216
        geon:98182">Baghdad</placeName></death>
</person>
```

network of authors

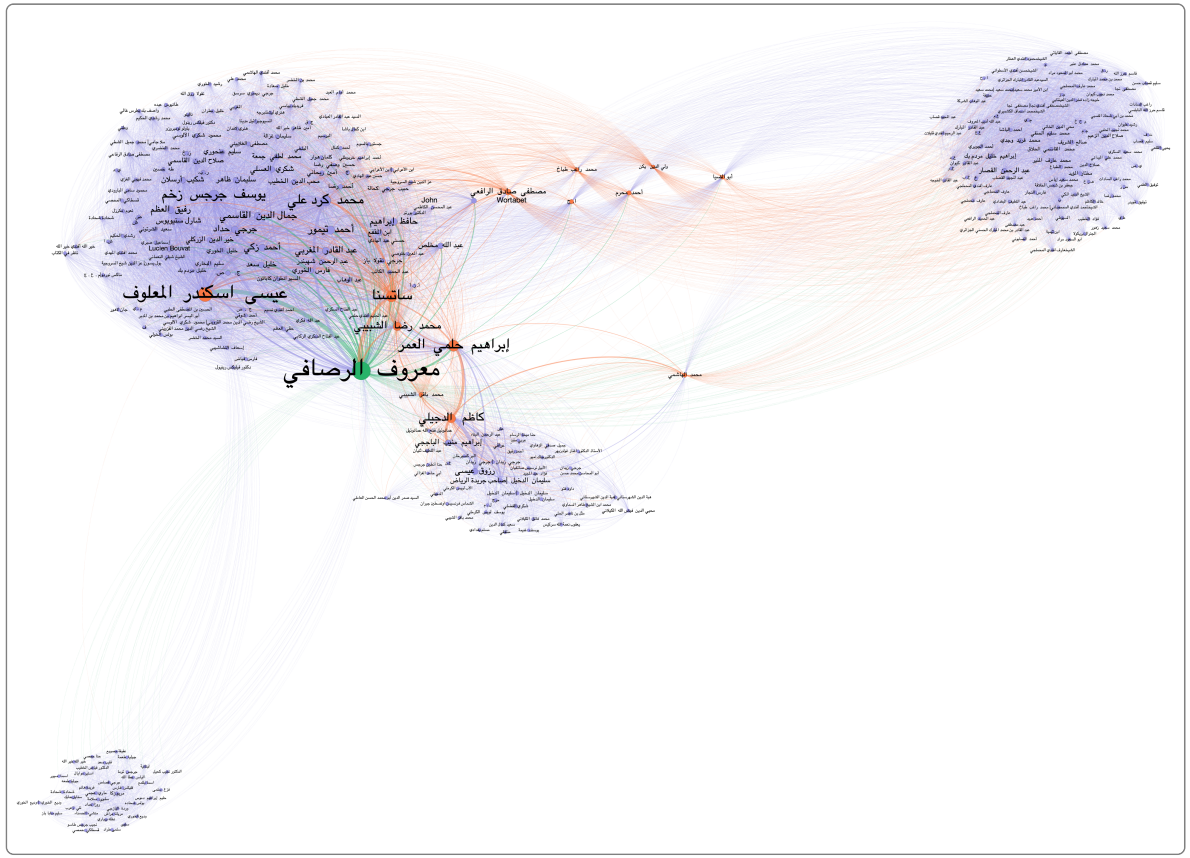


Figure: Network of authors with bylines in *al-Ḥaqā'iq*, *al-Ḥasnā'*, *Lughat al-'Arab* and *al-Muqtabas*

1. only a few nodes are of relative importance
2. limited overlap between journals from the same city
3. nodes are connected by various other social networks (education, imperial service, family relations etc.)

network of authors: the core

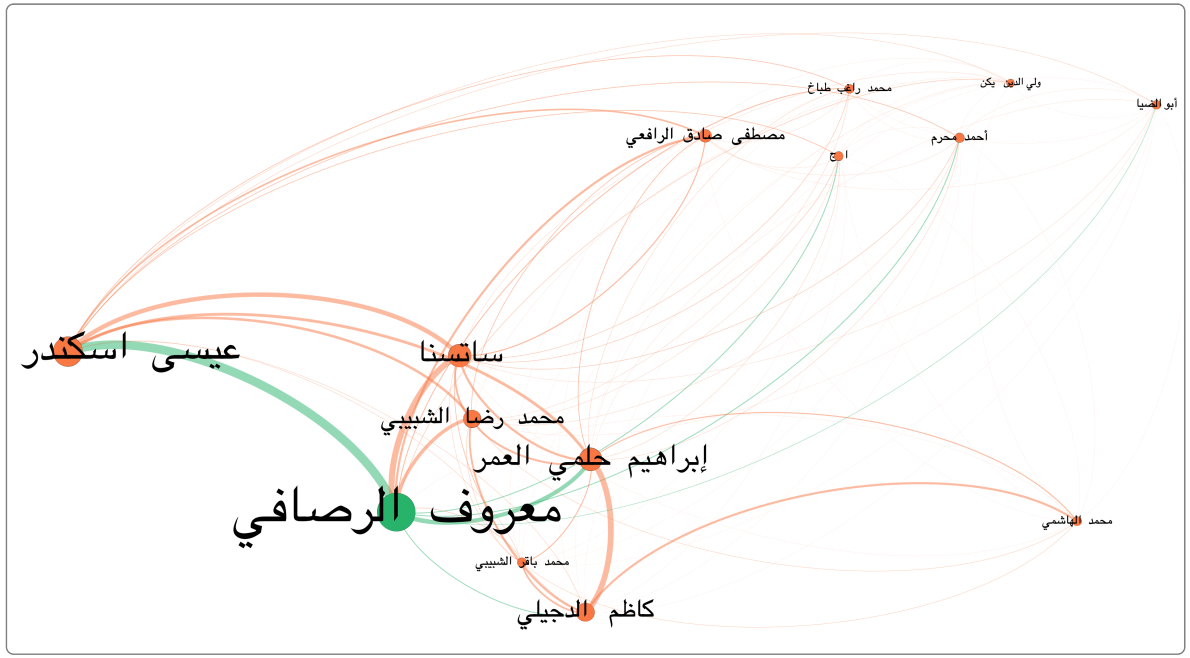
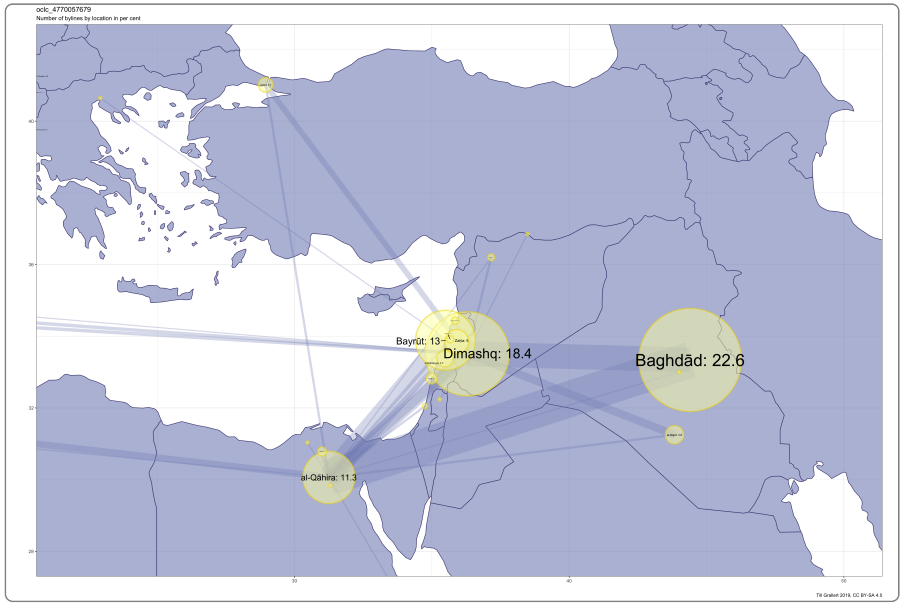


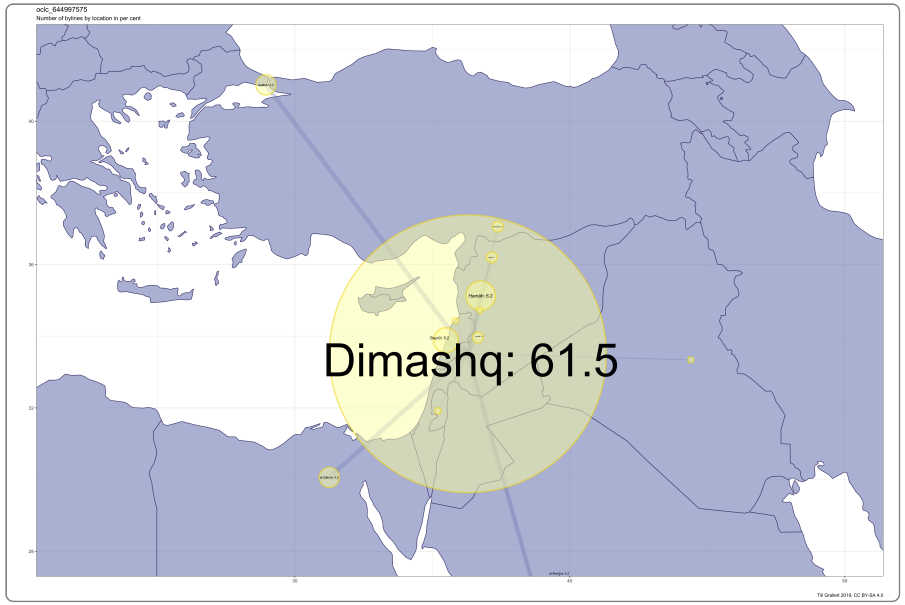
Figure: Core nodes in the network of authors with bylines in *al-Ḥaqā'iq*, *al-Ḥasnā'*, *Lughat al-'Arab* and *al-Muqtabas*

1. many Iraqis
2. few Syrians
3. few Christians
4. many poets
5. majority not covered by scholarly literature

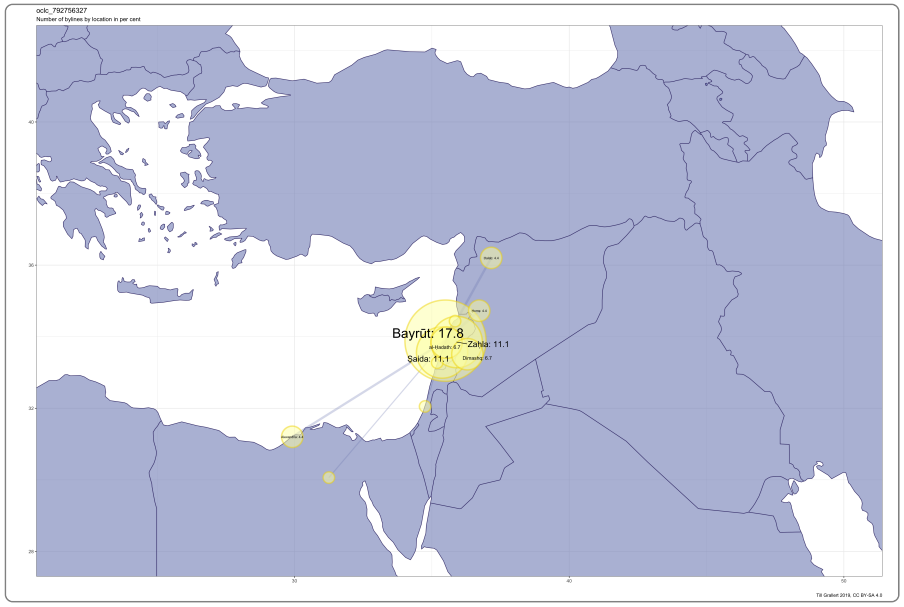
network of authors: geography



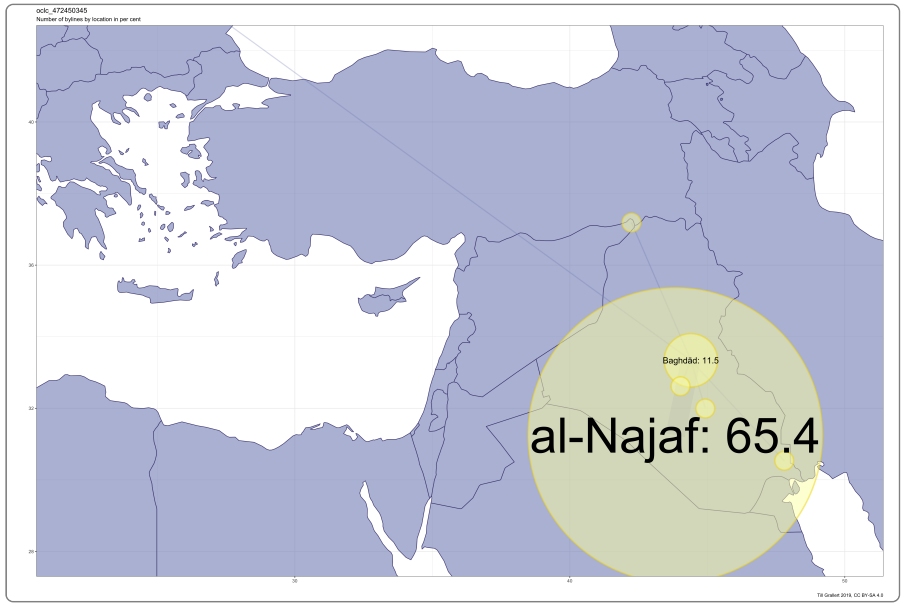
Map: Locations in bylines in *al-Muqtabas* (Cairo and Damascus)



Map: Locations in bylines in *al-Ḥaqā'iq* (Damascus)

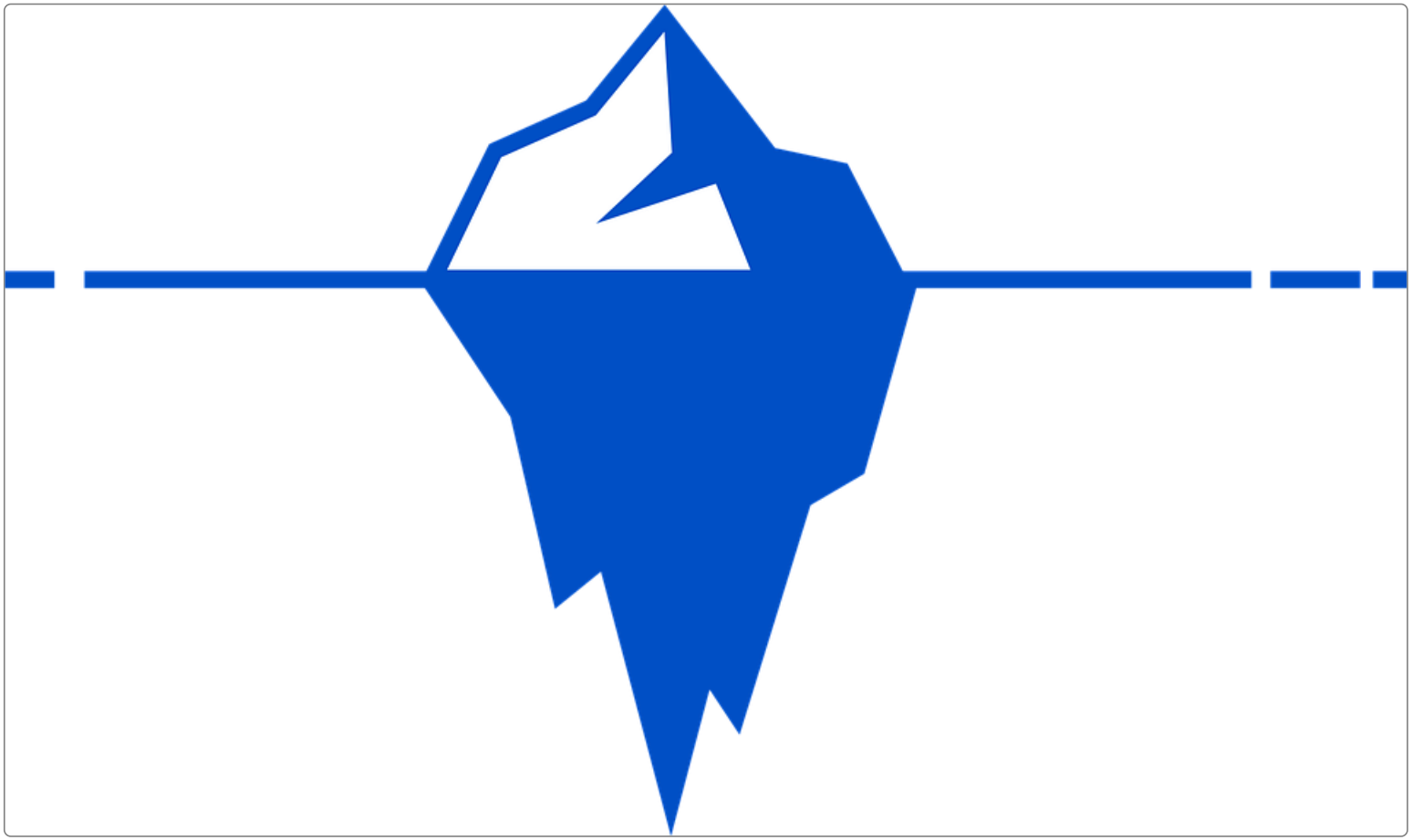


Map: Locations in bylines in *al-Ḥasnā'* (Beirut)



Map: Locations in bylines in *Lughat al-'Arab* (Baghdad)

problem: network reflects only 17% of articles



finally: computational authorship attribution

state of the field

- question of authorship attribution has not received much attention
- implicit and commonly accepted hypothesis: editors authored all anonymous articles themselves

problems:

- hypothesis remains untested
- we don’t know the potential candidates for authorship even within the hypothesis
- it is unlikely that a single person authored everything

computational authorship attribution: stylometry

- this has not yet been applied to Arabic
- method: *compare* stylistic features to gain a numerical measure of difference
    - distance measure is sensitive to composition of corpus
- stylistic features: most frequent words (MFWs)
    - have been shown to be extremely effective for authorship attribution
- chunking/sampling impacts results
    - features: run multiple iterations and have them vote
    - texts: minimum length of 4000-5000 words

stylometry: 5000 words

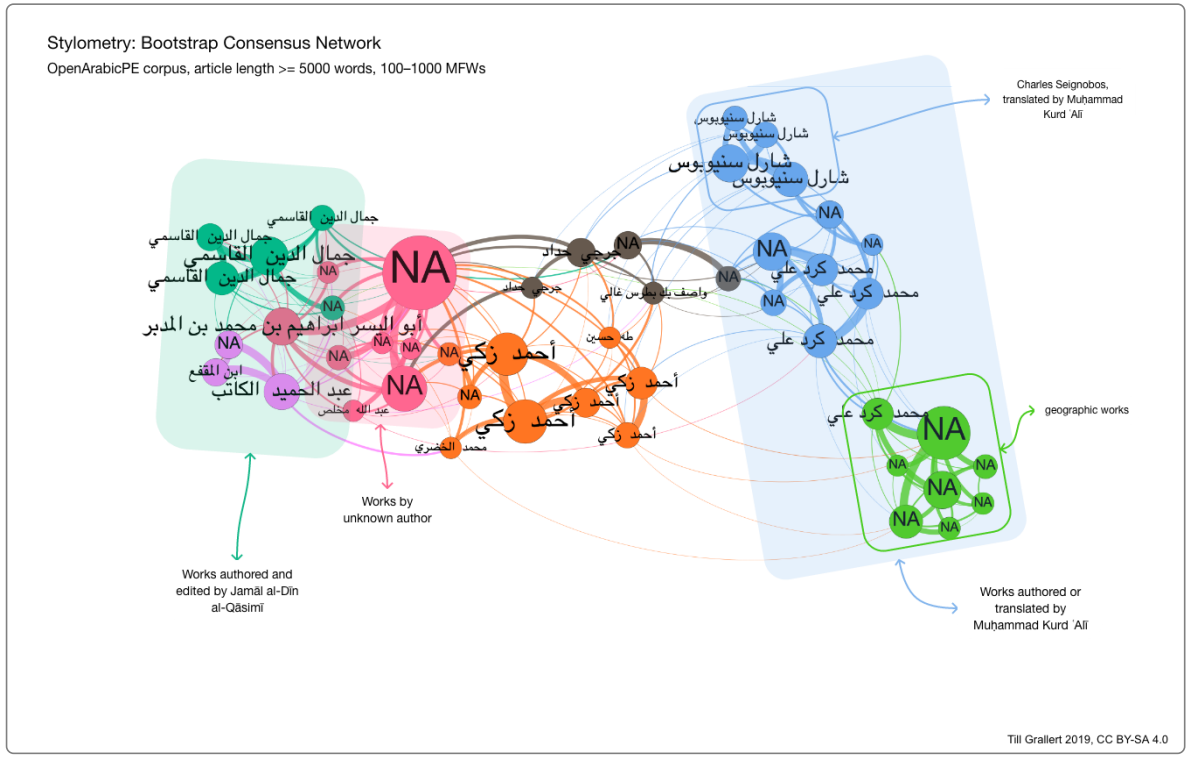
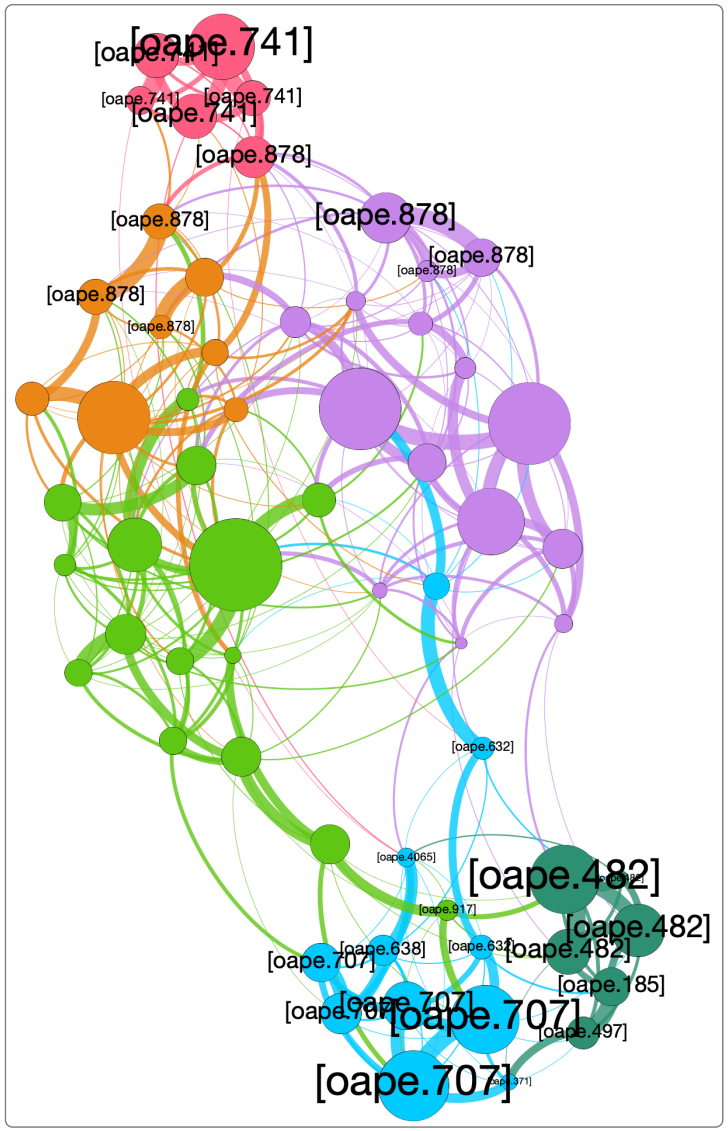


Figure: bootstrap consensus network of articles (length >= 5000 words, 100–1000 MFWs), colours by modularity group

- it works! correctly identified signal of
  - authorship
  - editorship
  - translation
- additional (sub)-signal: genre
- contradicts the hypothesis: anonymous author who is not the editor

stylometry: how to deal with the minimal length requirement?



- oape.878: Muḥammad Kurd ʿAlī
- oape.741: Charles Seignobos (in translation by Muḥammad Kurd ʿAlī)
- oape.482: Muḥammad Jamāl al-Dīn al-Qāsimī
- oape.707: Aḥmad Zakī Pasha
- oape.632: Jirjī Ḥaddād

Figure: bootstrap consensus network of sections and articles (length >= 5000 words, 100–1000 MFWs) in *al-Muqtabas*, coloured by modularity

stylometry: how to deal with the minimal length requirement?

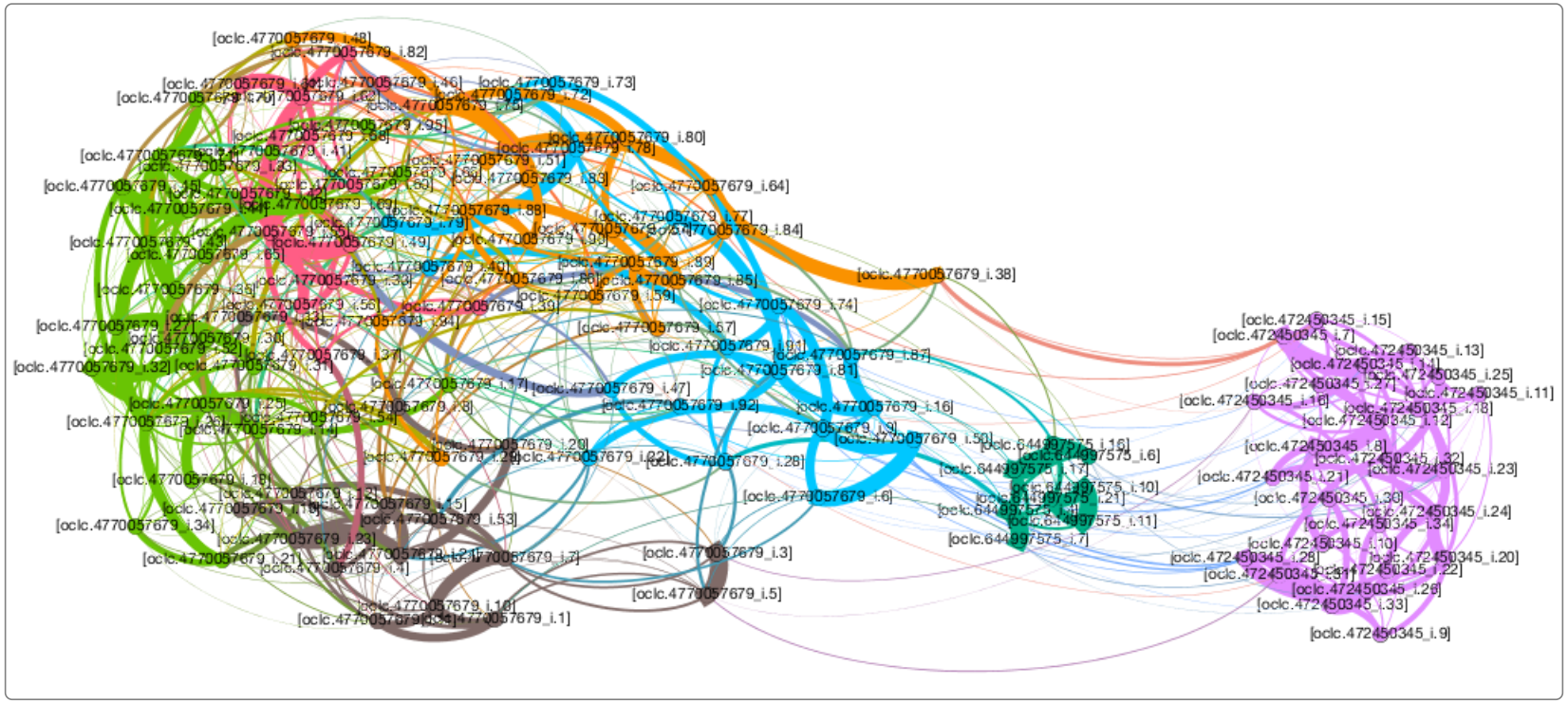


Figure: bootstrap consensus network of sections of articles in 123 periodical issues based on consensus of 100–1000 MWFs, coloured by modularity

Conclusion

Summary

- corpus building: modelling is indispensible for article-level analysis
- network of periodicals: confirms importance of Cairo
    - shows vast differences between journals: parochial, regional, trans-regional
    - should guide corpus building
- network of authors: surprising
    - importance of Iraq and Iraqis
    - very limited overlap between journals from the same city
    - many authors not mentioned in standard accounts
- stylometry: promising
    - problem: 5000 word minimum length

Thank you!

- Contributors to OpenArabicPE: Jasper Bernhofer, Dimitar Dragnev, Patrick Funk, Talha Güzel, Hans Magne Jaatun, Xaver Kretzschmar, Daniel Lloyd, Klara Mayer, Tobias Sick, Manzi Tanna-Händel and Layla Youssef
- Links:
  - Slides: https://OpenArabicPE.github.io/slides/2020-idh/
  - Paper (pre-print, submitted): https://doi.org/10.5281/zenodo.1413610
  - Project URL: https://www.github.com/OpenArabicPE
  - Project blog: https://openarabicpe.github.io
  - Twitter: @tillgrallert
  - Email: grallert@orient-institut.org
- Licence: slides and plots are licenced as CC BY-SA 4.0