## LDA2Net: Digging under the surface of COVID-19 topics in literature

Topic 50 companion sheet

G. Minello C.R.M.A. Santagiustina M. Warglien

This file contains the following supplementary information for Topic 50 of the manuscript "LDA2Net: Digging under the surface of COVID-19 topics in scientific literature":

- Human label and automatic n-gram label proposals (Table 1)
- Summary measures (Table 2)
- Network of top 25 bigrams (Figure 1)
- Wordclouds of top 25 words by node relevance measure (Figure 2)
- Wordclouds of top 25 bigrams by edge relevance measure (Figure 3)
- Filtered (0.99 percentile) topic network (Figure 4)

Table 1: Human and automatic label proposals. Automatic label candidate for largest word community of the topic. In parenthesis: absolute frequency of the walk out of a sample of size 1000.

Human label	2-gram label	3-gram label	4-gram label
NA	higher->risk $(13.6\%)$	adjusted->prevalence->ratio $(4.1\%)$	adjusted->prevalence->ratio->among $(2.3\%)$

Here follows the set of topic-specific measures that have been used to classify the topic and to analyse its structural properties (see manuscript for details):

Table 2: Summary measures

	JSD	Mean propensity	Variance propensity	Modularity	Barrat Clustering Coeff.
-	0.651162	0.008571	0.000309	0.304428	0.621818
	59	80	60	113	107

Based on the aforementioned measures, Topic 50 has been classified as a SPECIALIZED topic.

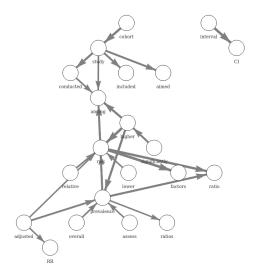


Figure 1: Network of top 25 bigrams (i.e., edges) by weight.

cohort included analysis higher RR age intervalconducted among ratio by study or prevalence in sortial association or prevalence or prevalence

used of conducted assess of higher aimed data prevalence padjusted risk CI resolution and association of analysis conducted assess of higher aimed data prevalence padjusted risk CI resolution and association of analysis analyses.

estimated participants ratios
ratios
polyconducted
study data
ippgprevalence
association riskration
age amongo

multivariable relative lower included lower interval lower included interval study by condition to the condition of the condi

RR conducted included respectively by the conducted respectively b

conducted participants
adjusted higher aimed association study data cohortratio risk CI total used among prevalence included factors analysis significantly respectively

Out-degree Betweenness PageRank

Figure 2: Top 25 unigrams (i.e., nodes) by measure.



Figure 3: Top 25 bigrams (i.e., edges) by measure.

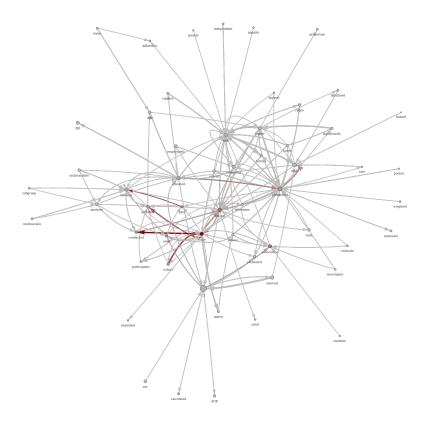


Figure 4: Filtered topic network (by weight). Layout based on Fruchterman-Reingold algorithm. Node size is proportional to topic-specific word probability provided by LDA. Edge width is proportional to topic-specific bigram weight provided by LDA2Net method. Node and edge color represent their betweenness centrality. Isolated nodes have been removed after filtration.