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

Topic 1 companion sheet

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

This file contains the following supplementary information for Topic 1 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
protective devices	protection->equipment (13.9%)	protection->equipment->used (5.5%)	filtration->protection->equipment->used (1.8%)

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.
value	0.523499	0.008676	0.000724	0.243170	0.521774
rank	9	82	107	109	11

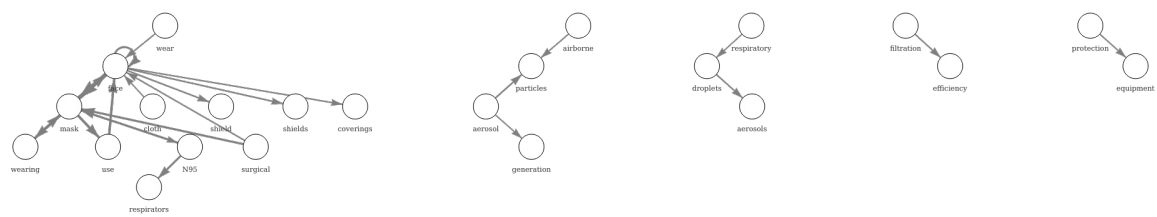Based on the aforementioned measures, Topic 1 has been classified as a SPECIALIZED topic.

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

LDA probability

exposure
protection
contamination
surfaces reduce
wearing particle
aerosol droplet
room disinfection droplets
face mask aerosols
used airborne
equipment use air
waste
particles N95
respirators filtration

Degree

respirator
shield contamination
airborne
protection respirators
aerosols aerosol wear
surfaces air face
disinfection mask
filtration wearing used droplets equipment
N95
surgical
particles droplet
efficiency

In-degree

disinfection
surgical equipment
efficiency surfaces
aerosol shields
shield use used droplets
droplet air mask
wearing respirator
particles N95
aerosols
protection
filtration
coverings

Out-degree

aerosols
equipment
droplets
respiratory surgical
protection
wearing mask wear respirators
face airborne
used air without
droplet aerosol particles
filtration cloth
contaminated reduce
respirator

Betweenness

contamination
droplet coverings
aerosol particles
respirators used
filtration generated mask respirator
aerosols face fit
virus use wearing
protection N95 equipment surgical
droplets covering
materials
spread

PageRank

spread
equipment
disinfection filtration efficiency
aerosols
generated aerosol droplets
surfaces via used use respirators
air mask airborne
face N95
wearing
particles protection
contamination droplet
surgical

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

Word cloud 1 (left):

respiratory>droplets
COVID-19>infection
filtration>efficiency
COVID-19>patients
healthcare>workers   mental>health
face>mask   patients>COVID-19
respiratory>syndrome   novel>coronavirus
infection>control   acute>respiratory
COVID-19>pandemic
coronavirus>disease
severe>acute   united>states
health>care   surgical>mask
public>health
N95>respirators
social>distancing
hydrogen>peroxide
spread>COVID-19
syndrome>coronavirus
care>workers

Word cloud 2 (middle):

contamination>face
mask>equipment
equipment>face
face>used   wearing>face
face>wearing   face>use
mask>use   use>face
mask>face   mask<>waste
face>face
face>mask   surfaces>face
mask>mask
face>aerosol
mask>wearing   use>mask   mask>aerosol
wearing>mask   used>face
particles>face   protection>face
mask>surfaces
face>protection

Word cloud 3 (right):

protection>equipment
mask>N95   filtration>efficiency
aerosol>particles
N95>respirators
airborne>particles   surgical>mask
wear>face
face>face   use>face
mask>face   wearing>face
face>mask
mask>use   cloth>face
face>shield   N95>mask
face>shields   wearing>mask
face>coverings   surgical>face
respiratory>droplets   droplets>aerosols
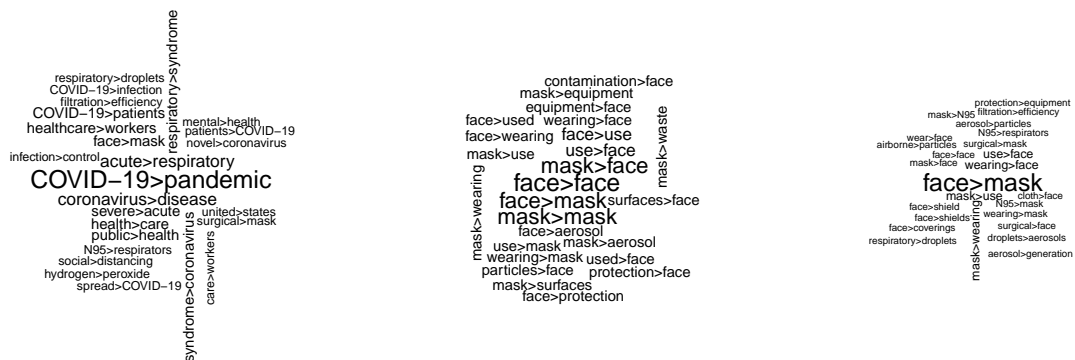mask>wearing   aerosol>generation

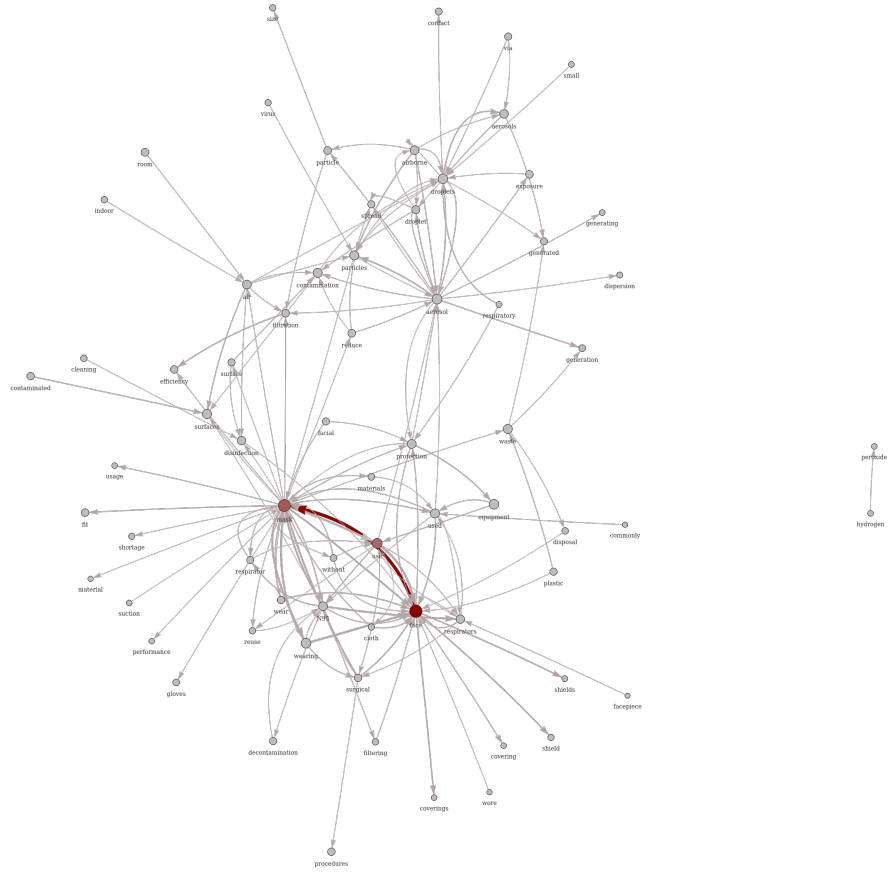Figure 3: Top 25 bigrams (i.e., edges) by measure.

4

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.