Harnessing the power of Digital Social Platforms to shake up makers and manufacturing entrepreneurs towards a European Open Manufacturing ecosystem

Deliverable Number	D2.4
Deliverable Title	Novel Network Metrics
Work Package Leader	BOUN

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 687941.

Grant Agreement N.	687941
Project Acronym	OpenMaker
Project Full Title	Harnessing the power of Digital Social Platforms to shake up makers and manufacturing entrepreneurs towards a European Open Manufacturing ecosystem
Instrument	Research and Innovation Action (RIA)
Thematic Priority	ICT–10–2015, Collective Awareness Platforms for Sustainability and Social Innovation
Start Date of Project	20 / 06 / 2016
Duration of Project	30 Months

Work Package N. | Title	WP2 | OpenMaker Digital Social Platform
Work Package Leader	BOUN
Deliverable N. | Title	D2.4 | Novel network metrics
Date of Delivery (Contractual)	Month 30
Date of Delivery (Submitted)	Month 31
Nature	Report
Dissemination Level	Public

VERSION LOG

Version	Issue date	Authors	Notes
0.1	08/12/2016	Xavier Sandin, YF	Document creation
0.2	13/12/2018	Angelo Facchini, Bülent Özel	First draft
1.0	21/01/2019	Angelo Facchini, Rossana Mastrandrea, Bülent Özel, Guido Caldarelli	Final

DISCLAMER

This document reflects only the authors' views and the European Commission is not responsible for any use that may be made of the information it contains.

EXECUTIVE SUMMARY

This deliverable presents a set of network driven metrics. The overall objective of these metrics is to provide machine learning supported recommendations to the enablers and members of OpenMaker community digital platform. The focus is on link formation dynamics between members where, depending on the context, the links can refer to online social ties and/or peer-to-peer channel of information flow between the members.

The approach for the metrics and the API services we present in this deliverable is contextual and data driven. State-of-the-art network theoretic approach is enriched by incorporating the individual's explicit or implicit interests, preferences, similarities and differences. The network metrics are further tailored according to the nature of the digital social medium and the modalities of the communication within the medium.

After providing the basic concepts of network theory, the deliverable describes a novel network metric based on similarity and difference among user members, leveraging on information declared by members themselves during the on-boarding phase in the OM DSP platform.

Network metrics describing the general topology of the OM networks are described, as well as the Network Analytics API, developed by IMT as a supporting tool for the Insight platform. The API provides the users and the community manager with tools, metrics and interactive maps to guide the user to the best exploitation of the OM community.

Finally, a recommendation system based on twitter data and network metrics is presented and discussed.

GLOSSARY OF TERMS

OM	Open Maker
DSP	Digital Social Platform
Explorer	DSP front-end user Interface for onboarded members
WatchTower	Backend data harvesting service of DSP
InSight	Backend analytics service of DSP
KPI	Key Performance Indicators
API	Application Programming Interface

Contents

1 Introduction

This deliverable presents a set of network driven metrics. The overall objective of these metrics is to be able to provide machine learning supported recommendations to the enablers and members of OpenMaker community digital platform. The focus is on link formation dynamics between members where, depending on the context, the links can refer to online social ties and/or peer-to-peer channel of information flow between the members.

The approach for the metrics and the API services we present in this deliverable is contextual and data driven. State-of-the-art network theoretic approach is enriched by incorporating the individual's explicit or implicit interests, preferences, similarities and differences. The network metrics are further tailored according to the nature of the digital social medium and the modalities of the communication within the medium.

2 Complex networks: a short introduction

In the recent years, thanks also to the access to large datasets, there has been an explosion of network models and analysis for the systems that are at the hearth of our society [DS14]. At the hearth of such models is the old an beautiful field of *graph theory*. The first paper of graph theory was written by Leonhard Euler and goes back by 1736 [Eul36]; however, the first textbook on graph theory is only in 1936, by Dénes Kőnig [Kon36].

Formally, a graph is an ordered pair $G = (V, E)$ where V is the set of vertices (also called nodes) and $E \subseteq V \times V$ is the set of edges (also called arcs or lines). Hence, to each edge $e \in E$ corresponds an ordered couple of vertices $(u, v) \in V \times V$. In the following, it will be considered the case of *undirected* graphs, i.e. $(u, v) \in E \to (v, u) \in E$; in such a case, an edge can be represented as an unordered pair of vertices $e = \{u, v\}$. Notice that in our notation it is impossible to have multiple edges, i.e. it is not considering the case of *multigraphs*.

The vertices belonging to an edge are called the ends or end vertices of the edge. A vertex may exist in a graph and not belong to an edge. The order of a graph is its number of vertices $|V|$; the size of a graph is its number of edges $|E|$. The degree $d(i)$ of a vertex i is the number of incident edges; self edges (i.e. edges of the form $\{v, v\}$) are counted twice.

A graph is called simple when it contains no multi–edges and no self–loops. A graph is complete if there exists one and only one edge between every pair of distinct nodes; is k–regular if all its nodes have the same degree (k) . An undirected graph is connected if every node can be reached from every other node. Finally, a graph $G = (V, E)$ is a subgraph of $G = (V, E)$ if $V \subseteq V$ and $E \subseteq E$.

A very convenient representation of a graph in terms of characteristic matrices associated with the graph. The most immediate representation of a graph G is its *adjacency matrix* A, i.e. a matrix whose ij^{th} element is 1 if there exists an edge between the $i^{th}$and the j^{th} vertices of G (fig.2). Notice that the degree of a node i can be defined in terms of the adjacency matrix as $d\left(i\right) = \sum_j A_{ij}$.

A powerful alternative for the matrix representation of the graph $G = (V, E)$ is given by its $|E| \times |V|$ *incidence matrix* B. To define B, let consider any the edges of G: let $e = (u, v)$ be the k^{th}edge, v the i^{th} vertex and u the j^{th} vertex with $i < j$. Then, $B_{ki} = 1$, $B_{kj} = -1$ and all the other elements of the k^{th} row are zero. Notice that the incidence matrix B is the network equivalent of the gradient operator ∇ in continuous spaces: given a vector $\{s_i\}$ of scalar quantities associated with the nodes, the difference $s_u - s_v$ of such scalar at the extremes of the k^{th}edge $e = (u, v)$ is $\sum_i B_{ki} s_i = s_u - s_v$, i.e.

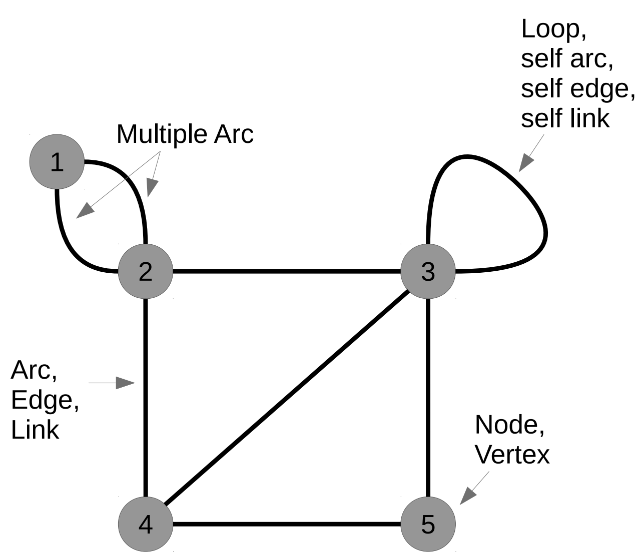


Figure 1: Graphical representation of an (undirected) graph. In the following, it will be considered simple graphs, i.e. graphs with no loops or multiple edges.

$$s_u - s_v = (Bs)_e \tag{1}$$

The incidence matrix B is very used in the engineering sciences to describe the topology of networks; by multiplying B by its transpose, another very important representation of a graph G is obtained, i.e. its *Laplacian* $L = B^T B$. Notice that, like in the continuous case where the Laplacian is defined as ∇^2, the network Laplacian is also defined as the "square" of the gradient operator B. By explicitly calculating the elements of $B^T B$, it can be seen that the $L_{ij} = -1$ if there is an edge between the $i^{th}$and the j^{th} vertices, L_{ij} is equal to the degree of the i^{th} node if $i = j$ and is $L_{ij} = 0$ otherwise. Hence, if it is defined by D the diagonal matrix whose ii^{th}element is equal to the degree $d(i)$ of the i^{th} node, it can also be written the Laplacian in terms of the degree matrix D and the adjacency matrix A as $L = D - A$.

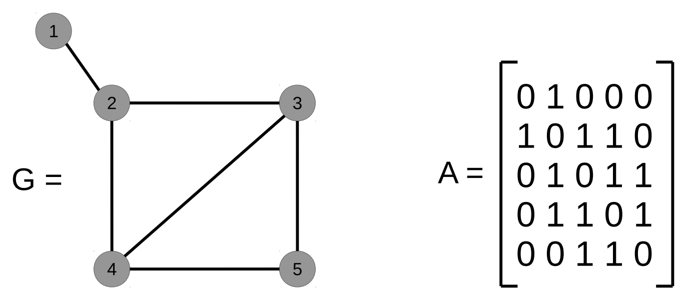


Figure 2: A graph G can be represented by its *adjacency matrix* A, i.e. a matrix whose ij^{th} element is 1 if there exists an edge between the $i^{th}$and the j^{th} vertices of G, 0 otherwise.

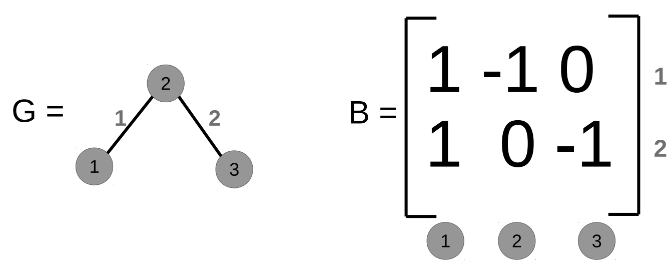
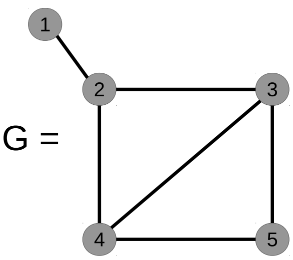


Figure 3: An alternative representation of a graph $G = (V, E)$ is given by its $|E| \times |V|$ *incidence matrix* B. Let $e = (u, v)$ be the k^{th}edge, v the i^{th} vertex and u the j^{th} vertex with $i < j$. Then, $B_{ki} = 1$, $B_{kj} = -1$ and all the other elements of the k^{th} row are zero.

Notice that the Laplacian matrix is often related to the dynamic properties of systems whose topology can be described as a graph [DSZC12].

G =



$$L = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 3 & -1 & -1 & 0 \\ 0 & -1 & 3 & -1 & -1 \\ 0 & -1 & -1 & 3 & -1 \\ 0 & 0 & -1 & -1 & 2 \end{bmatrix}$$

Figure 4: The ij^{th} element of the Laplacian matrix associated to a graph G is -1 if there is an edge between the $i^{th}$and the j^{th} vertices, it is equal to the degree of the i^{th} node if $i = j$ and is 0 otherwise.

2.1 Weighted Graphs

When graphs describe real systems, it is usual to associate quantities to the edges; in such a case, it is indicated of a *weighted graph*, i.e. a triplet $G = (V, E, W)$ where W is a set of quantities associated to the edges E. For an edge $e = (ij)$, let $w_e = w_{ij}$ the associated weight. The matrix representation of a weighted graph G is consequently modified: in the case of the adjacency matrix, the weighted adjacency matrix becomes

$$A_{ij} = \begin{cases} w_{ij} & \text{if } e = (i, j) \in E \\ 0 & \text{otherwise} \end{cases}$$

and the degree of a node becomes the sum of the weights $d(i) = \sum_j A_{ij}$ of the incident edges. Hence, the Laplacian matrix associated to a weighted graph keeps the form

$$L = D - A \tag{2}$$

where D is now the diagonal matrix whose ii^{th}element is equal to the weighted degree of the i^{th} node, i.e.

$$D_{ij} = \begin{cases} \sum_k w_{ik} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

.

In the incidence matrix representation, it is custom NOT to redefine the incidence matrix B, but to describe the system by the couple of matrices

(B, Y) where Y is a diagonal $|E| \times |E|$ matrix whose ee^{th}element is equal to the weight of the e^{th} edge, i.e.

$$Y_{ef} = \begin{cases} w_e & \text{if } e = f \\ 0 & \text{otherwise} \end{cases}$$

In the incidence matrix representation, the weighted Laplacian has the form

$$L = B^T Y B \tag{3}$$

2.2 Networks

While the word *graph* is associated to a very specific concept in mathematics linked to the very well assessed branch of graph theory, the word *network* has been come to assume a different nuance since the birth of of the so called *network science*. Network science takes birth just before the year 2000, first with the paper of Watts and Strogatz [WS98] and then with the paper of Barabasi and Albert [BA99]. Watts and Strogatz, to explain several apparently different systems, introduced a stochastic model for a class of graphs – Small World Networks – showing that by adding few random links it is possible to deeply change the properties of the network. Subsequently, Barabasi and Albert showed that a very simple mechanism of growth – preferential attachment – introduced another class of random graphs called Scale Free Networks that are characterized by a power law probability distribution of the degrees. Hence, year 2000 was the birth of *Complex Networks*, a field where Statistical Physics was applied to describe as statistical ensembles systems described by large datasets that could be mapped in a network.

3 Novel network metrics for the analysis of online communities

3.1 Network construction according to user skills

Given the set of members logged in the DSP, our need is to construct a network representation of their relations that keeps into account the following requirements:

1. Takes into account the peculiar characteristics of a community formed by inhomogeneous members, having widely different interests, skills, aims, and backgrounds.

Figure 5: Main skills present in the OpenMaker community, as declared by members

2. Maximise their probability of cooperation, especially fostering the maker-manufacturer cooperation.

3. Identify the top leaders in the community keeping into account the differences among the members, i.e. find a set of top influencers that are trusted by their reference community and at the same time are attractive for other members not working in the same field.

As pointed out in Deliverable 1.7, the OM community is composed by many persons, mainly makers, working in many different fields, as showed in figure 5. This also imply that in order to find a suitable and useful representation of possible connections among members we need to refer to a model based both on similarities and differences to define a possible connection among members. Therefore, when abstracting the DSP members into graph we developed a metric based on:

1. Tags[1]: as a proxy of the skills of each members.

2. Twitter: as a proxy of interaction among members.

Starting from Tags we have developed a metric based both similarity and difference. Starting from the Jaccard similarity to each pair of members A, B a measure of similarity is computed according to the following:

$$S_A = \frac{A \cup B}{A \cap B}, \tag{4}$$

where A and B are the tag sets of the two members, respectively[2]. while an asymmetric measure of difference has been implemented as follows:

$$D_A = \frac{A - (A \cap B)}{A \cup B}, \tag{5}$$

$$D_B = \frac{B - (A \cap B)}{A \cup B} \tag{6}$$

where D_A measures how A is different from B by A side, while D_B measures ho B is different from A by B side. To describe the potential link among two members we adopted the pair (S_A, D_B), using D_B to measure the difference from B side point of view. This choice is motivated by the fact that if A is interested in cooperating with B (therefore sharing a link), B should play a role in establishing such connection, i.e. if A is too much different from B side the link is not established.

A weighted network is then computed, and its matrix representation reads:

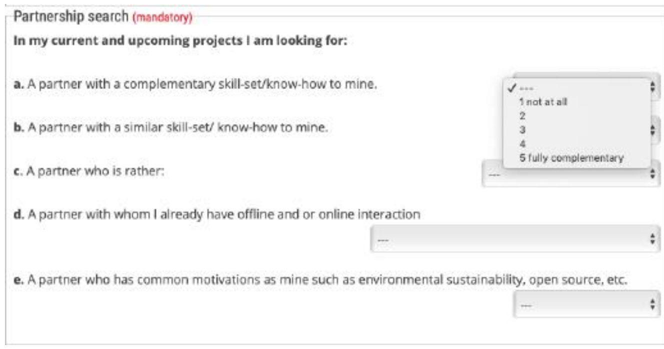
$$OM_{i,j} = (S_i, D_j) \tag{7}$$

In order to set a link between two members, a pair of thresholds is chosen according to the answers that the members have provided during the onboarding phase, where, a set of questions have been explicitly designed to provide insight about the willingness to cooperate with other members (see figure 6.

According to the answers collected, each member is characterised by two thresholds, ε_S and ε_D, and a link between two members (i, j) is established if both S_A and SD_B are above $\varepsilon_{S,D}$, respectively. It follows that the values

[1] Tags are set of keywords collected by members during the onboarding process. They are specific keywords describing the user's skills, beliefs, and main interests (see D1.7 for further details)

[2] By example, $A =$(3dprint, innovation, arduino, Fablab), while $B =$(Opensource, Education, Engineering, 3dprint, innovation).



Partnership search (mandatory)

In my current and upcoming projects I am looking for:

a. A partner with a complementary skill-set/know-how to mine.

✓ ---
1 not at all
2
3
4
5 fully complementary

b. A partner with a similar skill-set/ know-how to mine.

c. A partner who is rather:

d. A partner with whom I already have offline and or online interaction

e. A partner who has common motivations as mine such as environmental sustainability, open source, etc.

Figure 6: Specific set of questions aimed at understanding the willingness to cooperate with other members. $\varepsilon_{S,D}$ are extracted for each one of the members.

of $\varepsilon_{S,D}$ directly impact on the network, with different sensitivities, as shown in figures 7 and 8 where the reader can notice that the structure of the connection is mainly influenced by the value of ε_S. More in detail, figure 7 shows the average betweenness centrality of the network as $\varepsilon_{S,D}$ are changed, while figure 8 reports the average degree. As the reader can notice, the topology structure of the network is mainly affected by ε_S, while ε_D appears to be less involved in the link formation.

Another interesting feature of the network is that its representation is strictly centered on a specific user, because the values of $\varepsilon_{S,D}$ change from user to user. From this fact it follows that all the network metrics are centered on the user, and that the importance of the nodes based on centrality metrics (e.g. betweennes) are specifically tailored on the individual member, as shown in the maps presented in section 4 yielding a better performance of the recommendation methods that will be described in the following section 5, devoted to the recommendation systems.

3.2 Network metrics based on Twitter

In this section we present a set of novel community related metrics. The scope of these metrics are determined by the generation and propagation of the community related content on a popular social media. Data in public domain is curated by the software modules designed, developed and deployed for the project. Some of the metrics are used to generate a graphical user interface tool for the community enablers.

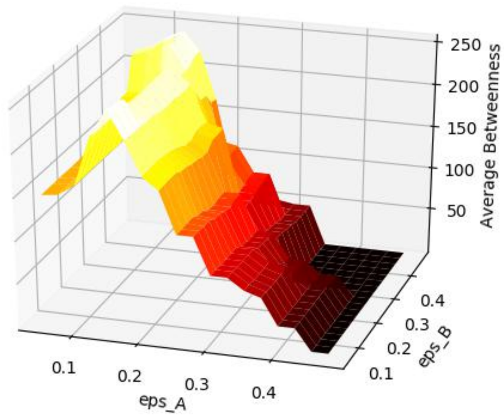


Figure 7: Average betweennes centrality of the nodes as $\varepsilon_{S,D}$ changes.

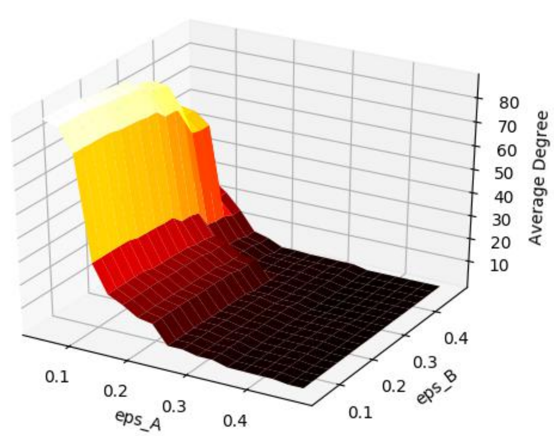


Figure 8: Average degree of the network as $\varepsilon_{S,D}$ are changed.

3.2.1 Discovering and locating leaders of community values: Community Spirometer API

The *community sprirometer* is software module that is designed, developed and deployed as an API as part of the OpenMaker Insight module. The deployment architecture of the API is presented in Figure 9.

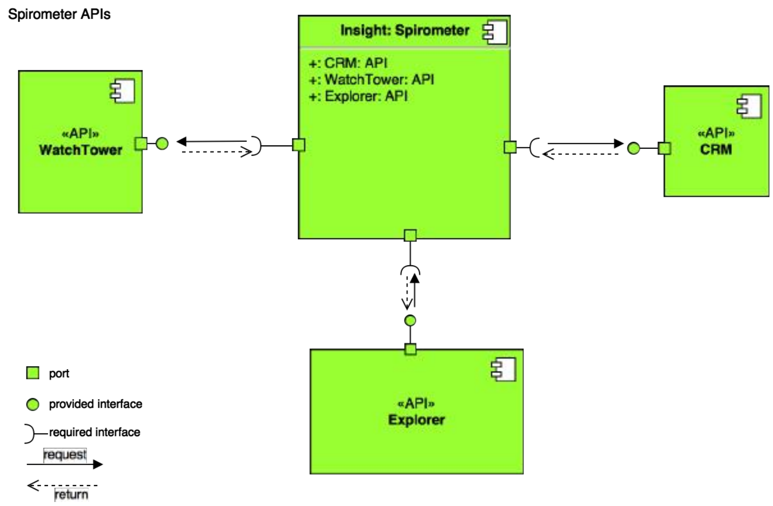


Figure 9: Architecture of the community spirometer.

The API provides an interactive graphical user interface to query and observe opinion leaders and influencers of the community. The focus of the module is to highlight those influencers who promote values of the open making or open making friendly social values.

The data is collected from the tweets that are in the public domain. The latest tweets in English of a community member are used for the analysis and visualizations, while the GUI version of the API is composed of two mutually interactive panels. The upper panel, see Figure 10, serves as member profiling as of his or her contribution to the community related debate. The lower panel, see Figure 11, is the community spiral on a selected dimension such as *Sustainability, Openness, etc.*

The nodes on the community spiral are resized according to the number tweets by the corresponding influencer. Depending on the user choice, influential actors can be placed rather in the core of the spiral giving a more visual readability for potential new members or at the outer branch giving a more direct readability on the leaders. In either case, the user is able to zoom in, zoom out or move along with the spiral inspecting details of each node such as number of tweets, score, type of score, and contribution level. The information is provided as a user hovers on a node. An influencer's spiral profile, as of his/her contribution to the maker movement related debates, is loaded into a separate panel when the node is tapped or clicked.

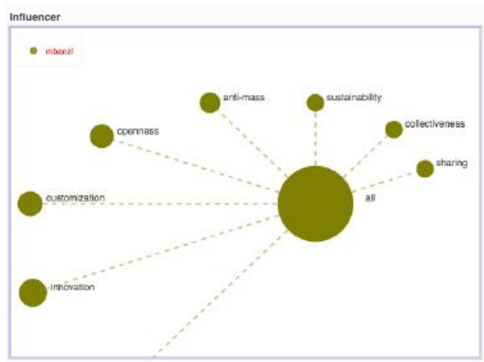


Figure 10: Contribution of an individuals to the community related themes.

Influencers of a given topic are clustered according to their sphere of influence. There are 4 or more distinct clusters on each spiral. Number of clusters and cluster membership is algorithmically generated from the data. A consequence of nodes with/without an inner circle denotes membership to the same cluster.

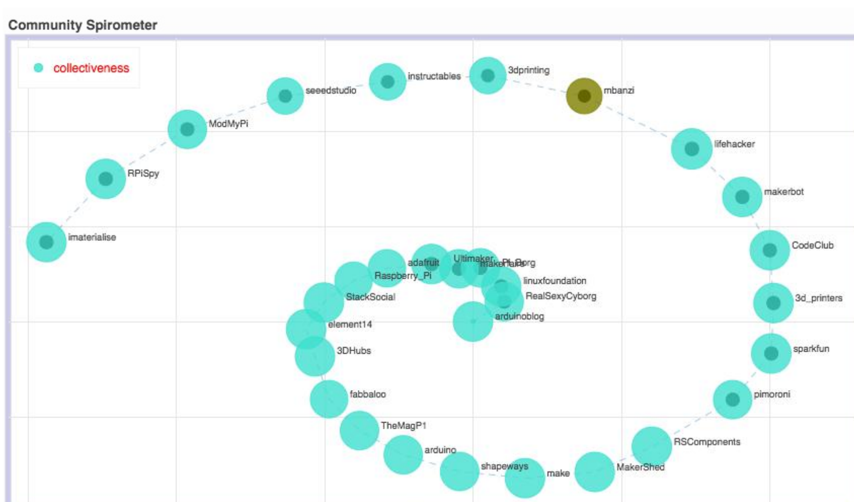


Figure 11: Community leadership.

Machine learning workflow of the API: The tool is based on a set of machine learning work flow that covers stages from collecting and labeling training data sets to automated scoring of input texts for their semantic content, and analyses of the community in terms of contributions of its individual members to the relevant debates via the public domain micro-blogging platform.

The **first stage** in the work flow has been identifying the themes and debates that are within the interest and focus of the maker movements in general. For this stage a desktop research has been conducted. A body of qualitative and survey based studies have been examined [Koh16, Ser14, Pai12, And12, Pow15, VH05]. Most of these studies are from the fields of cultural studies and socio-anthropological researches [Koh16, Ser14, Pai12]. The common approach at these mainly social science research is to consider the maker movement as a community of practice [And12, Pow15, VH05]. This very first stage allows for the identification of a set of concepts or issues that are recurring within the movement:

- Environmental **sustainability** through use of recyclable and local materials and resources,
- **openness** as of source codes of the software and the design of the hardware,
- **democratization** of product and access to the collective knowledge,
- **collectivism** in terms of collaboration and cooperation in processes of hardware or software productions,
- a free of cost at **sharing** and exchanging tools, source codes, and know-how,
- promotion and encouragement towards **innovations** in processes and products,
- ability and skills needed for a better **customization** of available tools, services and products,
- an **anti-mass consumption** stance that ignores individual needs, tastes and differences and solely motivate masses to consume more,
- a craftship skills that would enable individuals participate in processes of **making**.

In the **second stage** of the work flow, we have gathered and labeled texts needed to train the machine learning models. For each topic we have collected textual data where the primary topic of the text is the corresponding theme and participants of the debates are either practitioners or researchers in the area. Email lists, blogs, discussion forums, news, and articles were used. We have additionally employed a Wikipedia crawler specifically developed for the project in general to be able collect and label semantically related Wikipedia articles[3]. This stage has enabled us to form theme by theme specific corpora that are needed in the next stage. That is, for each one of the themes above we have created a collection of texts. Thus, for instance, a specific corpus on environmental sustainability is composed of the documents in which the primary topic is environmental sustainability.

In the **third stage**, we have cleaned and processed each corpus one by one in order to detect and learn the use of natural language and its choices terms and phrases. In other words, this stage has served our machine learning model at generating a likelihood score for a given input text to determine its relevance, for instance, to a debate on environmental sustainability. For the purpose, a separate software module is developed and published by an open source software distribution channel[4]. The first objective of this module is to provide a customizable and standardized text preprocessing prior to further analyses where more advanced machine learning and or statistical techniques can be applied and compared with each other. In that sense, it provides a pipelined set of functionalities (i) to be able to inspect, organize, prune and merge texts around one or very few specific theme(s) or topic(s), (ii) remove unwanted terms or literals from the texts, (iii) tokenize the texts, (iv) count the terms in texts, and (v) when desired stem the tokenized terms.

The second objective of this module is to be able compare or score a foreground corpus or a specific corpus against a background corpus or reference corpus. Example use cases could be, for instance, exploring the language of a sub-culture, a community, or a movement looking at to what extend the specific use of the language of the group differentiates itself from the common language.

At its specific use for the API, we have created ranked lists of terms and phrases for each maker movement related theme. It is assumed that terms and phrases we have extracted from a specific corpus tend to be relatively more frequent than a reference general corpus [SS15]. As of the general reference corpus we have used Brown Corpus. It is compiled by the researches at Brown University to be used as a present day general English corpus in the field of corpus linguistics. It contains 500 samples of

[3] See https://github.com/bulentozel/OpenMaker/tree/master/Scraping

[4] For the package distribution channel see: https://pypi.org/project/omterms/ and for the source code see https://github.com/bulentozel/omterms

English-language text, totaling roughly one million words [Mal02]. Given, for instance, a set of texts around open source software movement a term that is identified can be a word such as openness, a person such as Stallman a license type such as GNU, an acronym for an organization such as FSF the Free Software Foundation, or a technology such as Emacs.

We have devised a likelihood measure to compare frequency count of a term in a specific corpus versus its frequency count in the reference corpus. Here assumption is that the reference corpus is a large enough sample of the language at observing the occurrence of a term. Then having a higher/lower observation frequency of a term in the specific corpus is a proxy indicator for the term choice while having a debate on the topic. The likelihood ratio for a term P_t is calculated as:

$$P_t = \log \frac{ntS/NS}{ntR/NR} \tag{8}$$

where, ntS is the raw frequency count of the term in the entire specific corpus, ntR is the raw frequency count of the term in the reference corpus, NS is the total number of terms in the specific corpus, NR is the total number of terms in the reference corpus.

It should be noted that frequency counts are calculated after having applied the same tokenization and post processing such as excluding stop-words, punctuations, rare terms, etc both on the reference corpus and the specific corpus.

The **fourth stage** of work flow consists in the selection of a list of terms and phrases along with their content relevance scores to be used against external new input texts and implementation of the scoring algorithm. The likelihood scores of the equation 8 has been used as one of the inputs of the scoring process.

We have devised and implemented additional machine learning algorithms for the API. A k-means clustering algorithm has been implemented to identify phases on the community spirals and label members accordingly. Each algorithmically determined phase has been highlighted with a visual effect.

3.2.2 Bottom-up detection of topics within community debates

We have employed alternative tools at understanding online community dynamics in terms of the kind and types of debates they engage in. The methodology we have presented above is an hybrid approach where results from qualitative field research is fed in our supervised machine learning algorithms. Starting from a set of abstract concepts such as *sharing* we train our models to identify use of other terms, concepts, phrases and language around the concept, which in return helps us to discover community builders and influencers around the concept as exemplified by a sample community spirals in Figure 11. Additionally, we have devised a bottom-up methodology at capturing emerging and evolving trends within the community. The bottom up approach has a fully automated machine learning method.

We have employed a state of the art topic modelling approach that is based on the Nonnegative Matrix Factorization (NMF) [LS99]. The method decomposes a given nonnegative X matrix into W and H factors that are both nonnegative. More formally, given a $V \times T$ nonnegative matrix $X = \{x_{\nu,\tau}\}$ where $\nu = 1 : V, i = 1 : I$ and $\tau = 1 : T$, the objective is to find nonnegative factors W and H such that

$$x_{\nu,\tau} \approx [WH]_{\nu,\tau} = \sum_i w_{\nu,i} h_{i,\tau} \tag{9}$$

The approximate decomposition is obtained by solving the following minimization problem:

$$(W, H)^* = \arg \min_{W,H} D(X \parallel WH), \text{ subject to } W, H \geqslant 0 \tag{10}$$

In equation 10, the function D is a suitably chosen error function. The goal of using topic modeling is that to explore the hidden thematic structure of the documents, where the $V \times I$ matrix W captures distribution of documents to the topics, and $I \times T$ matrix H the relevance of of term to an underlying topic. In other words, while W is a representation of the document-topic relations, H is the representation topic-term relations. That is, the νth row of the W is the latent representation for document ν and τth column of the H shows the latent relationships of term τ with regard to I topics.

There are other popular topic modeling techniques such as Principal component analysis (PCA), vector quantization (VQ), latent Dirichlet allocation (LDA) and NMF. NMF is preferable since it provides more coherent topics[LS12, AC12].

In our model each tweet is considered as a document entry. Terms of each tweet is extracted using our *omterms* text analytic software package[5]. For the analyses tweets of the followers of the project is collected for a given

[5] See https://pypi.org/project/omterms/

period. The data is used to detect emerging topics. As of time of the analysis, Oct 3, 2018, there were 841 followers of the project. The top 3 emerging topics that our followers engage in are listed in Table 1.

Topic	Common terms
Policy making and maker movement	policy, maker, join, european, movement
Industrial innovation	solutions, new, innovative, innovation, website, hope, industrial
Event promotion	Rome, great, workshop, social, day, digital, ready

Table 1: Emerging topics as of OpenMaker Projects Twitter followers between January 1st to May 31st of 2018.

It is seen that the community is mostly engaged in a debate on policy making and maker movement related issues. Next they discuss and share about news, articles and thoughts on industrial innovation. It is also seen that the followers of the project are using the Twitter platform to promote maker related activities and events.

3.2.3 Alternative network metrics on social media influencers of a community: Twitter based metrics

We have designed a set of network driven metrics from a community formation perspective. The metrics are based on the online interactions both in terms of content generation and sharing, as well as, as of following-follower, mention and retweet links. The metrics we have elicited or proposed have a common objective. We aim to identify community leaders, community related content generators, community related content propagators, and embeddedness within the community. Besides we have devised a set of simple network metrics where the objective is the identify potential new members of the community. These metrics combined with stylized network metrics such as *centrality, betweenness, closeness*, etc. can extend insights on community dynamics from a network theoretic perspective.

The approach is applied to the public domain online data of Twitter platform. Twitter is essentially a micro-blogging platform where the size of each blog entry is limited by the character length of the text. However, a blog entry may contain a URL giving pointers to other contents. Content generated or received from other users can be re-published. Special characters can be used to send notified references to the other accounts within the platform.

For the clarity, it should be noted that the boundary of the community is drawn based on the following definition:

Definition 3.1. OpenMaker Twitter Community The OpenMaker Twitter community is defined by the twitter accounts that are following OpenMaker's Twitter account, which is namely **@openmaker**.

3.2.4 Network Types

Given rich public domain meta data that can be accessed via Twitter's API we are able to construct three types of networks as laid out below. Each node in these networks corresponds to a public account on the micro-blogging platform.

Definition 3.2. Who-is-followed-by-whom Network It is a directed unweighted graph. The existence of the link denotes a subscription relation. An incoming directed edge from a destination node to the source node denotes direction of information propagation where the source node is the information generator and the destination node is the subscriber. In other words, the followers of an account are the incoming edges of the corresponding node on *Who-is-followed-by-whom Network*.

An extended analysis on part of this network is presented in Section 4.2, where of all the followers of the community only on-boarded members as of our CRM records are considered at the formation of the network.

Definition 3.3. Who-mentions-whom Network It is a directed and weighted graph. The existence of the link denotes a mention relation. The source node on a directed edge is the community member who mentions the destination node. The weight of the link expresses the number of mentions within the time-window used to construct the network.

Definition 3.4. Who-retweets-whom Network It is a directed and weighted graph. The existence of the link denotes a retweet relation. An incoming directed edge from a destination node to the source node denotes direction of information propagation. The source node is either the original node of the tweet or an intermediary node. The weight of the link expresses the number of retweets within the time-window used to construct the network.

3.2.5 Metrics

RT popularity: The measure stands for retweet popularity of a member for a given time window. It is a reputation score representing the number of times a follower's post is retweeted by the community.

$$Pop_i^{RT}(t_0, t_f) = \alpha^{RT} \frac{\sum_{t_0}^{t_f} n_i^{RT}(t)}{\sum_j^N \sum_{t_0}^{t_f} n_j^{RT}(t)} \tag{11}$$

It should be noted that t_0 and t_f specify a time window as of calendar days and $n_i^{RT}(t)$ is the observation function for the number of tweets by community member i at a certain day t. The multiplier α^{RT} is an arbitrary scaling factor for human readability; when it is set as $\alpha^{RT} = 100$ it would point to a percentage share of the member. The equation normalizes the retweet popularity with respect to the total volumes of retweets generated by the community.

Alternatively, the measure can be computed from *who-retweets-whom* network that is constructed from the data set. Then for a given node i the normalized total weight of its outgoing edges, eij would capture the node's *RT popularity*:

$$Pop_i^{RT} = \alpha^{RT} \frac{\sum_j e_{ij}}{\sum_i \sum_j e_{ij}}. \tag{12}$$

Mentioned popularity: The measure captures the popularity of a member within a community as of number of times he or she has been mentioned by the other members. Similar to the metric above, we normalize the measure as of the total number of mentions within the community and for a given time window:

$$Pop_i^{@} = \alpha^{@} \frac{\sum_j e_{ji}}{\sum_i \sum_j e_{ji}}. \tag{13}$$

Note that from a network theoretic perspective the measure is the summation of incoming edge weights of a node within *who-mentions-whom* graph.

Knowledge generators: The measure aims to capture a member's behavior in terms of content generation. We have two alternatives for this metric. The first one examines the node at the ego level comparing the ratio of original content respect to the total number of micro-blog posts by the member herself/himself. The second one compares the type of the post with respect to the total posts by the community. A Twitter post that is not a retweet is considered as an original post:

$$K_i^{org,ego}(t_0, t_f) = 1 - \sum_{t_0}^{t_f} \frac{n_i^{RT}(t)}{n_i^{T}(t)}. \tag{14}$$

Knowledge propagators: The metric highlights information propagators within the community. It captures the members in the node who more inclined to help diffusion of information by retweeting a blog post within the platform. The behaviour can be considered a factor at collaborative knowledge filtering, where a repeatedly retweeted content is assumed to point out a higher relevance or interest within the community. It is a complementary to the measure that is described above:

$$K_i^{dif,ego}(t_0, t_f) = \sum_{t_0}^{t_f} \frac{n_i^{RT}(t)}{n_i^{T}(t)}. \tag{15}$$

$$K_i^{dif,Net} = \alpha^{dif} \frac{n_i^{RT}}{\sum_j^N n_j^{RT}}. \tag{16}$$

It should be noted that N in 16 represents the total number of members in the community and hence the metric is a ranking as of total volume of retweets within the community. On the other hand, 15 can be considered as a characterization of the member as of his or her content generation behaviour.

Community cohesiveness: We have devised a set of metrics in order to measure the relation of a member to the community. For those measures we examine node properties on the networks that we have constructed. As of type of relation, we focus on inclusiveness, cohesion and embeddedness. In other words, we aim to measure association of a member to the community in terms of its position on the network as well as his or her interaction with the other members of the community via the means of information sharing on a popular public sphere social media.

In these measures we basically compare preferences of the user as of its interaction with the other member of the community vs non-member of the community.

Reciprocity at Who-is-followed-by-whom Network: We have devised two simple network metric which measures the level of reciprocity within the community. From a network theoretic approach the measure is the ratio of reciprocated edges between pairs of nodes to the total number of edges in the corresponding directed graph:

$$Reciprocity = \frac{\sum_i \sum_j e_{ij} e_{ji}}{\sum_i \sum_j e_{ij}} \tag{17}$$

It should be noted that $e_{ij} = 1$ if there is a link from node i to node j, otherwise $e_{ij} = 0$. Thus the measure above is a network level aggregate measure hinting a level of cohesion within the community as a whole.

Cohesiveness at Who-mentions-whom Network, $Co_i^{@}$: For a given node we compute the number of times the node mentions a member of the community versus his or her total mentions. The ratio, $0 \leq Co_i^{@} \leq 1$, would have a maximum value for a member who mentions only the other followers of the community account. It would be minimal for a member who mentions other members of the micro-blogging platform that are not yet followers of the community account. The measure can be used as a proxy to capture to what extent the community is within the focus of the user. A cumulative statistics over all the followers may give an idea of the cohesion between followers.

It should be noted $1 - Co_i^{@}$ can be interpreted as a proxy for the social capital of the member as a gateway to the new members.

Cohesiveness at Who-retweets-whom Network, Co_i^{RT}: For a given node we compute the number of times the node retweets a content from another member of the community versus his or her total retweets.

In a similar manner to the mention related cohesiveness, the complement of the measure: $1 - Co_i^{RT}$ can be considered as an indicator for the potential of the member as an information gateway for the community.

Node vs network level measures: It should be noted that all the metrics we have presented in this section are node level measures for a given snapshot of the network. Community level aggregate versions of them can be generated in a straightforward manner. Such aggregate versions can be traced as of moving time windows to be able to observe evolution of the network. Besides, these measures can be employed to compare and contrast different communities or the snapshots of the same community at its different point on its life cycle.

4 Analysis of the OpenMaker network

In this section we describe the OM network as it is analysed by means of the `Network Analytics API`, available through the `Insight` server. The OM network is composed by about 375 members, and considering an average of the values $\varepsilon_{S,D}$ is described by the metrics presented in table 2. In the following we will consider an average of $\varepsilon_{S,D} = 0.4, 0.6$, in order to provide the reader with a snapshot of the whole community[6].

Number of Explorer members	375
Number of links	2172
Network diameter	8
Network betweenness centrality	0.148

Table 2: OM Network

Referring to Table 2, the number of links provides a measure of the level of potential cooperation that is present in the community. In this specific case, the number of connections indicate that the density of the network, computed by means of:

$$D = \frac{2(E - N + 1)}{N(N-3) + 2} \tag{18}$$

is about 2.57%, a value in line with a community made of heterogeneous members. With regards to the network diameter, a value of 8 means that the shortest distance between the two most distant nodes is 8, therefore a maximum of 8 steps are needed to put in connection each member of the community. The higher the diameter the harder is to reach a member far in the network. The above mentioned topological characteristics of the networks find a further confirmation in the average betweenness centrality: 0.148 is a low average value, indicating that many members share few connections with others, so the OM networks is expected to be mainly composed by subgraphs of people connected by few hubs, i.e. nodes that have a significative number of connections, as we can see looking at figure 12 showing the map of the OpenMaker community according to the values of $\varepsilon_S = 0.6$ and $\varepsilon_D = 0.4$. Each member is represented by a blue circle localised in the position declared during the onboarding phase, while the diameter of the circle is proportional to the betweennes centrality of the node: the larger size, the higher importance of the node in the network. By means of this representation the user has access, at glance, to the top influencers of his own network (meaning that other users with different thresholds usually see different influencers). The map is interactive, zoom is enabled and the name of the member can be accessed by clicking on the circle.

Going more into details, each member has also access to the map showing all the other members sharing with him a connection, as depicted in Figure

[6]Descriptive statistics of the community have been extensively presented in Deliverable D1.7.

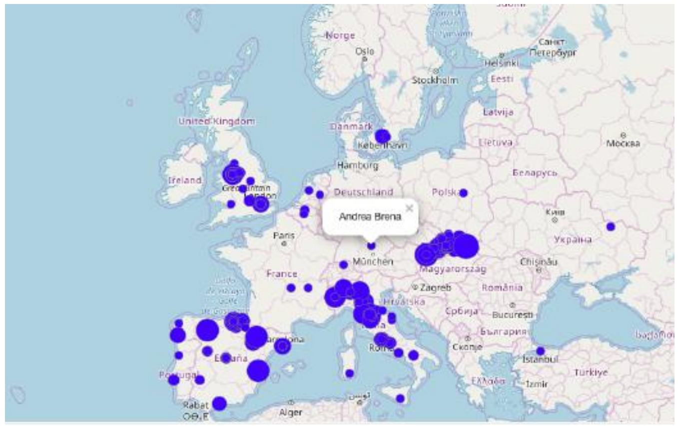


Figure 12: Map of the members of the OpenMaker community. Radius is proportional to betweennes centrality. The map is interactive: the user can click on a member's spot to find information.

13, where the yellow circle represents the user and the blue circles represent the location and influence of the other members directly connected to him. Again, the map is interactive, and the name of the member can be accessed by clicking on it.

4.1 Community detection

Given the maps and the network structure showed in the previous section, it is useful both for the user and for the community manager understanding if inside the community there are set of members that share the same tags or is there exists a number of subsets (or subgraphs) composed by people sharing a common set of skills or beliefs. To unfold the community and to better enter into details, a community detection method has been implemented to search for graph subsets in the DSP community.

In general, a network is said to have community structure if the nodes of the network can be easily grouped into (sometimes overlapping) sets of nodes such that each set of nodes is densely connected internally, i.e. such group of nodes share more connection with respect to other nodes of the network. A community structure divides the network into groups of nodes with dense connections internally and sparser connections between groups. But overlapping communities are also allowed. A general definition is based on the principle that pairs of nodes are more likely to be connected if they are both members of the same community, and less likely to be connected if

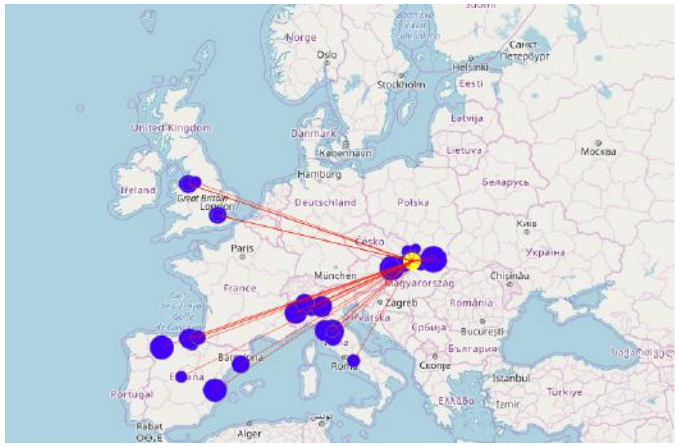


Figure 13: User centered network: Each member have access to a personalized map showing his connections to the other members.

they do not share communities.

Finding communities is important because it sheds light on hidden patterns and connections not immediately visible by inspecting the network (especially in case of wide social networks), as communities often correspond to functional units of the system. Furthermore, being able to identify substructures within a network can provide insight into how network function and topology affect each other.

Existence of communities also generally affects various processes like innovation diffusion or spreading of ideas and messages happening on a network. Hence to properly understand such processes, it is important to detect communities and also to study how they affect the spreading processes in various settings.

To detect communities in the OM network, we considered the Louvain algorithm, based on the concept of optimal modularity and well known for its ability to cope with low density networks (for further details about the Louvain method the reader is referred to BIB).

Figure 14 shows the community map as it is produced by the `Network Analytics API`. As for the previous case, the map is fully interactive, and shows with different colors the communities that are automatically detected in the DSP network. Each community is organised in layers, so the user can add or remove layers and find influencers guided by the size of the spot.

Each community is characterized by a set of tags, such characterisation has been discussed in Deliverable D1.7, while in figure 15 we report the

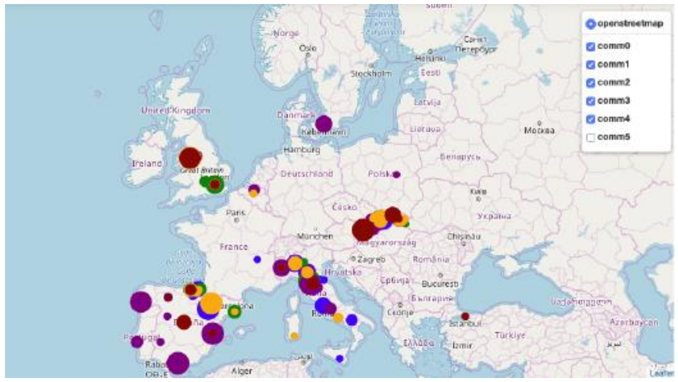


Figure 14: Detected communities in the OpenMaker network. Radius is proportional to betweennes centrality; The map is interactive, and the user can select the communities of interest having access to the names of the members.

findings for the sake of clarity.

The `API` also provides an user centered network showing the members of his own community, as showed in figure 16, where the user is depicted in yellow, the radius of the circles is proportional to betweennes centrality, and the connections among the members are in red.

4.2 Trust metrics based on Twitter

In this section we study the network of Twitter followers for the CRM users who provided their Twitter account name. Indeed, this represents a sub-network of the global network of all OM users. Those data have been downloaded using the API developed by the team of Bosphorus University: it allows to obtain for each user the list of all its Twitter followers and to select, among them, only the DSP members. The resulting network is directed and has 158 nodes and 739 links.

If we neglect link direction, the network is composed by 6 disconnected components: 6 groups of nodes internally connected, but whose between-connections are absent. Specifically, there is a dominant component of 148 nodes and 5 groups of two nodes each. We decided to focus on this component only with the aim to avoid meaningless outcomes biased by the other minor components. The novel Twitter network has 148 nodes and 728 links, resulting very sparse with a density ~ 0.03 (see Figure 17).

Taking into account links directionality, which constraints the paths of

Community “ Innovation Designers” (73 members)		Community “Software innovators” (62 members)		Community “3D makers” (14 members)		Community “Creative Designers #2” (77 members)	
Design	66	Innovation	37	3dprint	6	Design	60
Innovation	59	Software	23	Arduino	4	Creative	21
Sustainability	15	Development	13	3d scanning	4	Maker	16
3d	9	Entrepreneur	12	Raspberrypi	3	Opensource	14
Enterpreneurship	8	Iot	10	Innovation	2	Social innovation	13
Community 4 (14 members)		**Community 5 (4 members)**		**Community 6 (5 members)**		**Community “Researchers” (94 members)**	
Innovation	10	Additive	2	Community	4	Innovation	67
Collaborator	5	Slicer	2	Empowering	3	Research	30
Artist	5	Science	2	Opensource	3	Engineering	28
Fablab	4	Metal	2	Enabling	2	Education	21
Enterpreneurship	4	Programming	2	Prototype	1	Creativity	17

Figure 15: Most common tags in the communities detected by the Louvain method.

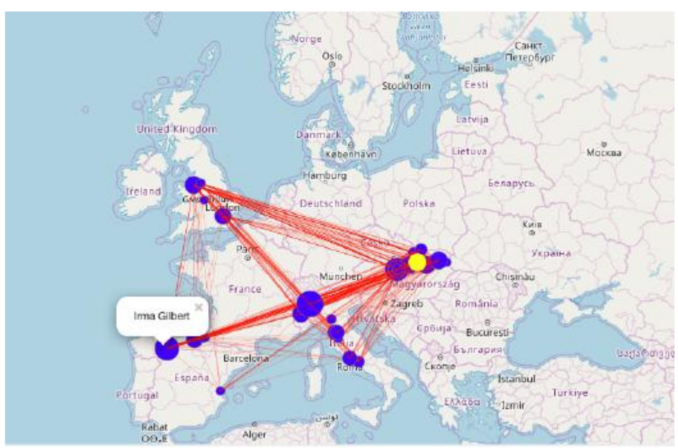


Figure 16: Belonging community of the single member. The user is depicted in yellow, radius is proportional to betweennes centrality, map is interactive.

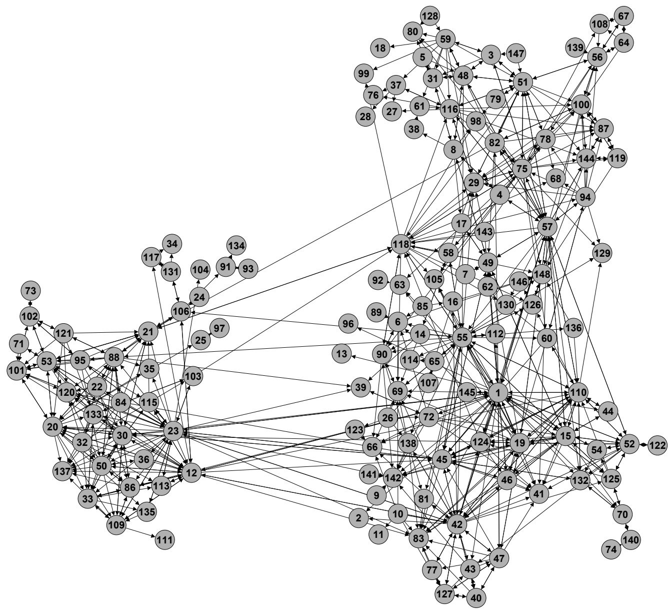


Figure 17: The Twitter network of CRM users (giant component). The link directionality represents the relation 'who-follows-who'. Node labels are attributed to the list of CRM user names in increasing order (see Table).

the information flux, the network consists of 47 components. This means that, according to the relation who-follows-who, the network is formed by 47 groups of nodes that are connected in one direction but not in the other. In other term, following the link directionality is not possible to cover all the network. In Figure 18 there is an example of a network made by 1 weakly connected component (ignoring link direction), but 3 strongly connected components.

This is a preliminary information that starts to shed light on the architecture of the Twitter network of CRM users. Indeed, it reveals a general absence of reciprocal link and a tendency to have node with high number of both incoming links (in-degree) and outcoming ones (out-degree). This is evident in the scatter plot between the two quantities (see figure 19) and their significant correlation value equal to 0.67. On the other hand, nodes between these very connected CRM users have mainly edges in only one direction creating constraints in the networks flow paths.

Node degree represents a first simple measure of node centrality, as it reveals node importance at local level, just looking at first neighbours. In

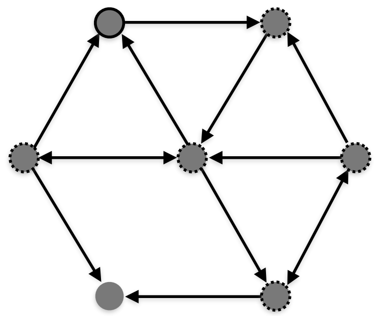


Figure 18: An example of weakly and strongly connected components of a network: nodes with same border-line representations belong to the same components.

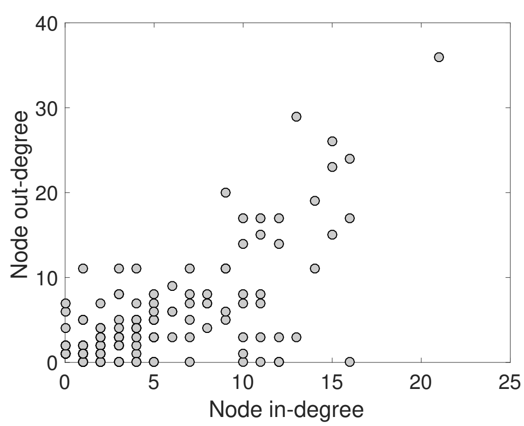


Figure 19: Scatter plot between the node in and out degree distributions of the Twitter network of CRM user followers.

directed network with several strongly connected components, a more informative way to measure node centrality consists in looking at its hub and authority ranks, measure that assigns a centrality score to a node looking at the centrality of its neighbours. Both measures are strictly interconnected and can be associated to each node. Generally speaking, authorities are nodes which give important information on an interesting topic; while hubs are nodes telling where there are the more relevant authorities [New18]. In this context it means that a node in the top hub ranking follows nodes that have high authority, i.e., a large number of followers themselves, and so on (see fig. 20 for an example).

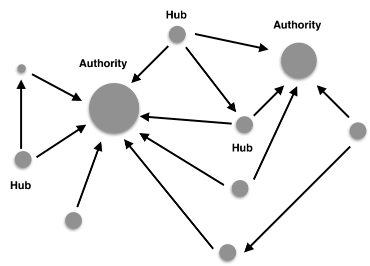


Figure 20: Example of nodes with high hub and authority scores in a small toy-model network.

The idea has been introduced by Kleinberg [Kle99] who also developed an algorithm called *Hyperlink-Induced Topic Search* (HITS). In Kleinberg's approach the hub centrality of a node is assumed to be proportional to the auhtority centrality of all nodes it points to; while the authority centrality of a node is proportional to the sum of the hub centrality of all nodes pointing towards it. Therefore, they are computed in a recursive way for all nodes in the network updating the hub and authority rank of a node looking at the whole chain of its connections (neighbours of neighbours of neighbours and so on).

Formally,

$$\begin{aligned} h_i &= \alpha \sum_{i->j} a_j = \alpha \sum_{j=1}^{N} A_{ji} a_j, \quad \forall i = 1, ..., N \\ a_i &= \beta \sum_{i<-j} h_j = \beta \sum_{j=1}^{N} A_{ij} h_j \quad \forall i = 1, ..., N \end{aligned} \tag{19}$$

where, α, β are constants and A represents the adjacency matrix of the network.

From (19) is clear that a high hub centrality score of a node is a combined effect of the number of nodes it points to and their authority centrality weights; on the other hand, a high authority centrality score of a node is a combined effect of the number of nodes pointing towards it and their hub centrality weights.

We can express the (19) also in matrix form as follows.

If we consider the vectors of the hub and authority centrality scores of all nodes, $\overline{h}$, $\overline{a}$, we can write:

$$
\begin{aligned}
\overline{h} &= \alpha A \overline{a} \\
\overline{a} &= \beta A^T \overline{h}
\end{aligned}
\tag{20}
$$

Equations in (20) can be substituted into one another, giving:

$$
\begin{aligned}
\overline{h} &= \lambda A^T A \overline{h} \\
\overline{a} &= \lambda A A^T \overline{a}
\end{aligned}
\tag{21}
$$

where $\lambda = (\alpha\beta)^{-1}$.

Thus the hub and authority centrality are respectively given by the eigenvectors of $A^T A$ and AA^T associated with the same eigenvalue, and this is exactly the leading eigenvalue [New18]. It is worth to notice that bot relations in (21) hold if $A^T A$ and AA^T have the same leading eigenvalue λ. However, it is easy to prove that both matrices have the same eigenvalues [New18], and in particular the laeding one.

Figure 21: The Twitter network of CRM users (giant component). The link directionality represents the relation 'who-follows-who'. The node dimension is proportional to the Hub centrality score.

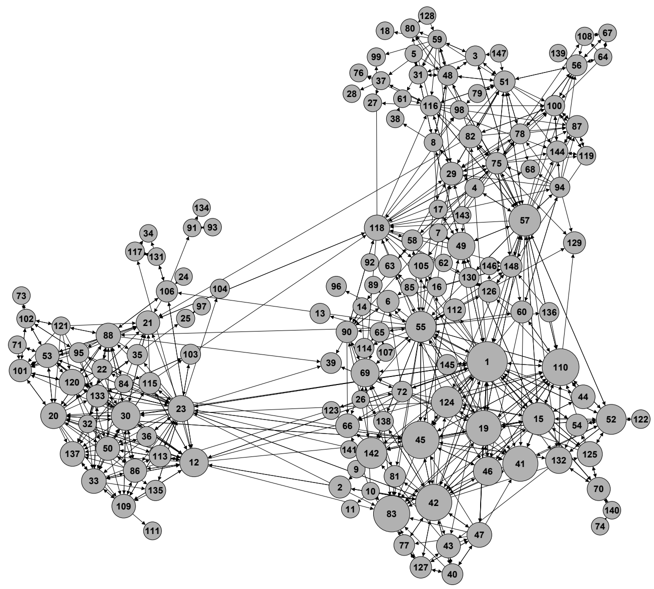


Figure 22: The Twitter network of CRM users (giant component). The link directionality represents the relation ‘who-follows-who’. The node dimension is proportional to the Authority centrality score.

In figures 21 and 22 we report the Twitter network of followers of CRM users in 17 with node size proportional to the hub and authority centrality scores, respectively. Two main features are evident: (i) there are few nodes with very high hub and authority centrality scores; (ii) nodes tend to have similar hub and authority centrality scores. Indeed, there is high significant correlation $\tilde{0}.73$ between the two distributions of scores (fig.23).

There are few exceptions,like nodes $41, 69, 83$ exhibiting higher authority centrality than hub one. Therefore, we can say that in general nodes follow nodes having several other followers that are in turn hubs, while are followed by nodes tending to follow many other nodes that are in turn authorithies and so on (see figure 20 for a clear example). This outcome represents a specific feature of this network characterized by few very central nodes who follows and are followed by a number of CRM users, while most of the nodes just have few links in one of the two directions. Hence, nodes in top 5 or top 10 hub and authority centrality ranking represent a good recommendation for other CRM users as they follow several authorities and are followed by a number of hubs; in other terms, they could have a fundamental role in the spreading of information, innovation, knowledge on the global network

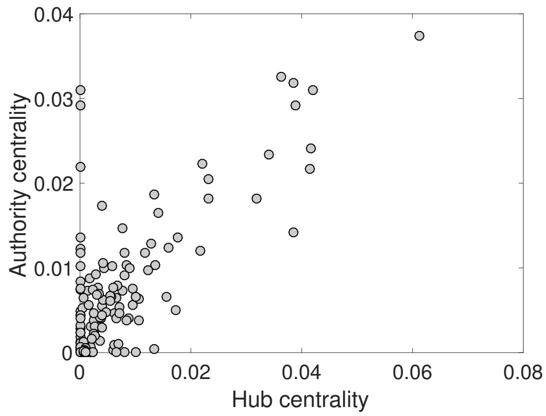


Figure 23: The Twitter network of CRM users (giant component). The link directionality represents the relation 'who-follows-who'. The node dimension is proportional to the Authority centrality score.

of CRM users.

5 A recommendation system based on network analysis and trust

In this section, we provide the conceptual design of a recommendation engine that is designed and implemented for the platform. The design of the engine aims to produce personalized recommendations for the members of the platform. The conceptual approach of this novel recommendation engine has been already applied at implementation of a personalized network analytics API that we have presented in Section 3.1 and Section 4.

5.1 A conceptual approach to similarities vs differences

We consider that members of the digital platform may look for new partners not only for their commonality in expertise and social capital but also for the differences.

The types of recommended items are listed as follows:

- technology of interest,
- skills of interest,
- project partner,
- project topic.

The personalized aspect of the engine stems from the fact that the recommendations are tailored to following member specific input sets:

- preferences,
- skills,
- area of interests,
- social network connections,
- community affiliations,
- geographical locations.

Our novel approach is based on actual data sets. That is the machine learning models behind the recommendations are trained by live data. Observed skills, technologies as well as project specifications within the community are used to derive compatibility and similarity scores between pairs of members as of their area of interests. Their mutual and distinct social network connections are used to measure their affinity to collaborate. Their explicitly declared preferences regarding their collaboration behaviours are used both at ranking and filtering their potential collaborators. The data driven technology and skill ontology is used at recommending new skills or technologies to acquire.

The on-boarding form of the digital social platform is organized to be able to collect user inputs that can be used to train the modelling components of the recommendation engine. An additional survey has been devised to standardize and validate data sources for our machine learning modules. For automated data schema and data validation an independent software package has been designed, implemented and added to the open source software distribution channels[7]. For the the collection of survey data two complementary methods have been devised. The first one adopts online survey tools where members of maker communities are invited to complete the form. The second more novel approach employs the chatbot application designed and implemented for the DSP. The survey is an extended version of the on-boarding form. Thus the chatbot is designed to complete the form interacting with the user at his/her later visits to the site. Instead of linear and a single session interaction the survey questions are presented at different times. The selection of the questions is tailored to the activity of the member according to the page content he or she is navigating through.

5.2 Modelling components

In order to train underlining models for the recommendations two sets of data are used. The first data set is constructed from maker projects. Technologies and skills used or required for a project is extracted and transformed into technology-technology, skill-skill and skill-technology co-occurance data sets. The second data set is constructed from maker profiles. This second data set enables us to construct bases of abstract profiling of individuals and sub-communities. It should be noted that although the two data sets are correlated inferences that can be derived from them are distinct. A maker's profile on skills, education, or use of technology may contain ontologically irrelevant items, whereas skills or technologies declared on a project serves as the basis of compatibility of skills and technologies.

These data sets are represented in relation matrix formats such as project by skills, project by technology, profile by skill and profile by technology.

[7]For package distribution see https://pypi.org/project/omdata/ and for the source code see https://github.com/bulentozel/omdata

Then for instance, a maker's capacity to transfer an existing skill to the required set of skills for a new project partnership are derived employing ontological and semantic relations captured by these mathematical relations.

In other words, such indirect relations are derived in an organic manner from existing data sets. Skills and technologies that are identified at past projects are used to capture moving and evolving frontline of skills and the transferability of know-how from one skill to another or from previously popular skill sets to emerging new ones. We implement a set of advanced machine learning algorithms to derive indirect relations between skill sets. Underlying factors and singular value decomposition of observed skills are computationally identified to be able to generate distance metrics in our normalized mathematical model of skill space. Vectoral proximity in our skill-space is used, for instance, at catching compatibilities of different profiles.

In addition to direct and indirect relations between a hypothetical project and a maker profile, we also develop a scoring scheme to evaluate fitness of a profile and the impact of longevity of past experiences. While variety and consistency of the skill portfolio of a maker is measured for his/her capacity to adopt agile development processes. A non-linear diminishing return function is used at determining impact of years of experience in an area.

5.3 Recommendation based on communities and centrality measures

The `Network analytics API` provides a recommendation system based on the following methods:

1. Top influencers of the global network, suggesting the members whose score is in the top 10 of all members (see by example figure 12).

2. Top influencers in the member's community, suggesting the top 5 members having the highest rank (see by example figure 16).

3. Top influencers of the other communities, suggesting for each community the top 5 members (see by example figure 14).

4. Top influencers of member's network (see by example figure 13)

The number of top influencers is fully configurable, and can be restricted or increased by the community manager.

5.4 Example: A recommendation network as of area of interests

In this example we assume that all of our members have the same type of behaviour:

- **Reciprocity:** They collaborate with some other member only if they mutually agree to set-up a collaboration.
- **Homophily:** They all have the same minimum required cognitive similarity threshold. They think the level of commonality in their expertise will enable them to be able to communicate and work together.
- **Heterophily:** They all have the same minimum required cognitive difference from each other. They seek certain level of differing expertise to be interested in a collaboration.

For the presentation we have used self-appointed tags of OM members. The output of our model is a recommendation network where each edge in the network imply a potential partnerships between nodes (peers).

Figure 24 presents the surface model for the number of recommendations. The surface is model can be considered as a way to explore potential partnerships. It is derived from self-declared tags. Note that x-axis denotes the level of difference between members whilst y-axis the level of similarity, and z-axis is the potential size of collaboration for a given level of similarity and difference.

The potential number of partnerships in this model are determined by varying the required level differences and similarities. These levels are assumptions on the thresholds of the preferences of individuals; symmetry in preferences are assumed. The point where the similarity and the difference is minimal, i.e. equals to zero, the number of recommendations would simply indicate the total number of indirect links that can be established between nodes of a given network.

The figure further suggests that:

1. Given initial expertise distribution of our members, the relation between similarity threshold and potential number of partnerships is nonlinear: With an increasing cognitive similarity threshold, we first observe a sharp decrease in potential partnerships and then we start to observe new potential partnerships between very similar members. This, in a way, demonstrates the existence of significantly very similar expertise profiles on our platform. An increasing cognitive difference

threshold has rather more linear impact on the potential partnerships. However, it never goes down to zero. This implies that an expertise seeking behaviour can foster and sustain partnerships between our members.

2. We have a community-wide profile where combined effect of modest cognitive similarities and differences can foster sustainability of the partnerships. The corresponding area is marked with a label in Figure 24.

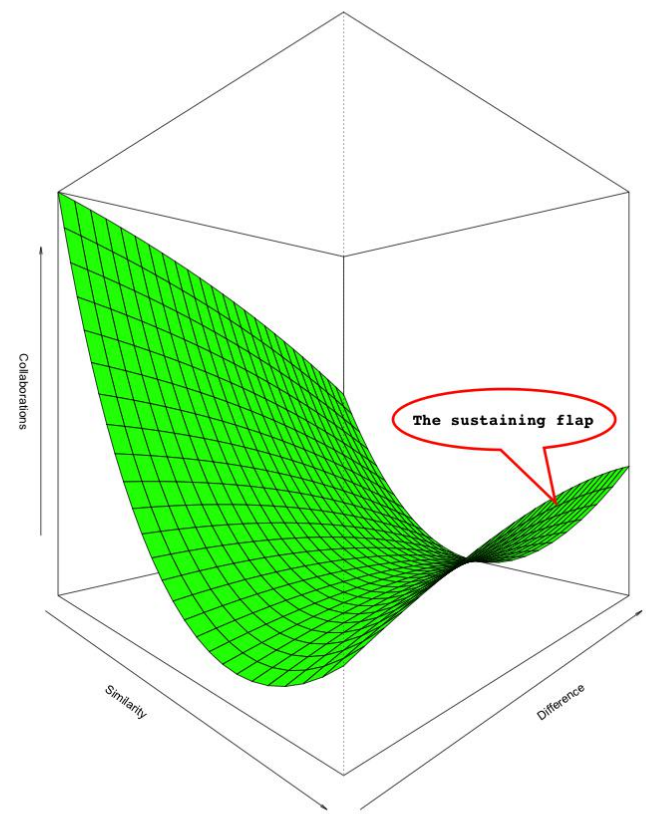


Figure 24: A surface model for the potential number of collaborations between OM community members as of their are of interests.

In our generic approach a recommendation takes into account both similarities and differences. Our conjecture is that a productive partnership may arise when members with complementary skills get together. In other words, there must be at least a sufficient level of similarity to be able to communicate yet there must also exist differences in expertise to attract and complement each other for a creative process.

Figures 25, 26, and 27 present networks of recommendations for partnerships between the members of the OM community where a minimum similarity threshold of 0.25 and a minimum difference threshold of 0.75 are assumed. It should be noted that these thresholds correspond to a specific point on the 3D surface model of Figure 24. More precisely, we have generated a network representation of the potential links for the point we have labeled in the figure above.

Figure 26 is representation of a potential partnership network for the whole OM community. The direction of a link indicates who is recommended to whom. The nodes are colored according to their geographical locations at the country level. In this recommendation network it is assumed that all of the members are indifferent to the location of another member in the community at their willingness to initiate a partnership. The uniformly observed reciprocation as of directed edges in this particular network graph is due to homogeneity assumptions – as of difference and similarity in areas of interest as well as at geographic proximity – in the preferences of the individuals. In reality asymmetries arise due to heterogeneous preferences.

Figure 26 presents a potential network of international partnerships between the members of the OM community, while Figure 27 is a potential network of local partnerships. Nodes in the figures are colored according to the country level geographic locations. Both networks are essentially different sub-graphs of the recommendation network drawn in Figure 24. In these sub-graphs the preferences as of geographic location of a potential collaborator are updated. Namely, while in Figure 26 it is assumed that the OM members consider only international collaborations for the given threshold levels required for the cognitive similarity and difference, in Figure 27 it is assumed that the OM members only seek for local collaborators with the same levels of thresholds.

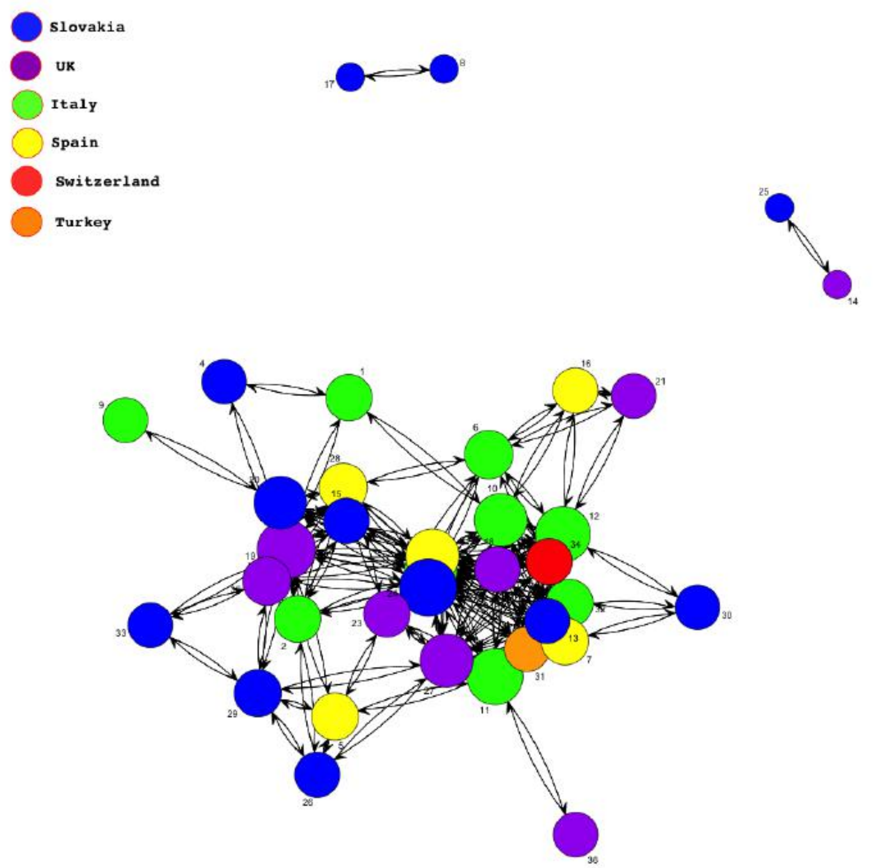


Figure 25: A potential partnership network for the whole OM community. The direction of a link indicates who is recommended to whom. The nodes are colored according to their geographical locations at the country level. In this recommendation network it is assumed that all of the members are indifferent to the location of another member in the community at their willingness to initiate a partnership. Besides it is assumed that each one of them seeks a 0.25 level of similarity and a 0.75 level of difference at a potential collaboration. The uniformly observed reciprocation as of directed edges in this particular network graph is due to such homogeneity assumption in the preferences of the individuals. In reality asymmetries arise due to heterogeneous preferences.

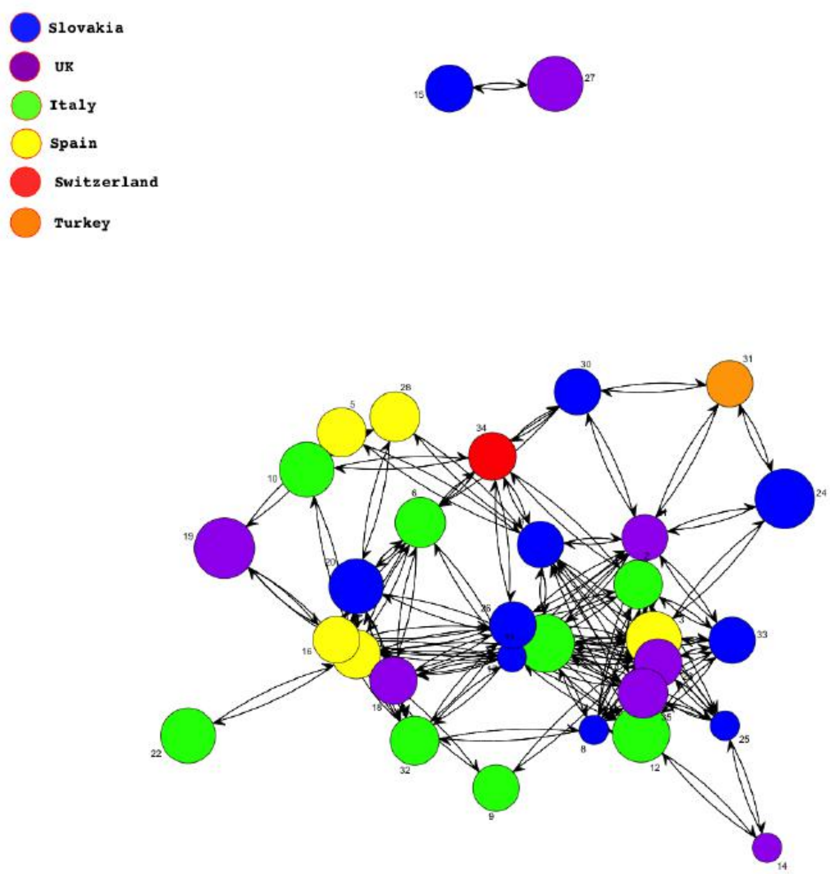


Figure 26: A potential network of international partnerships between the members of the OM community. Nodes are colored according to the country level geographic locations. This network is essentially a sub-graph of the recommendation network drawn in Figure 24 where only the preferences as of geographic location of a potential collaborator is updated. Namely, here it is assumed that OM members consider only international collaborations for the given threshold levels required for the cognitive similarity and difference.

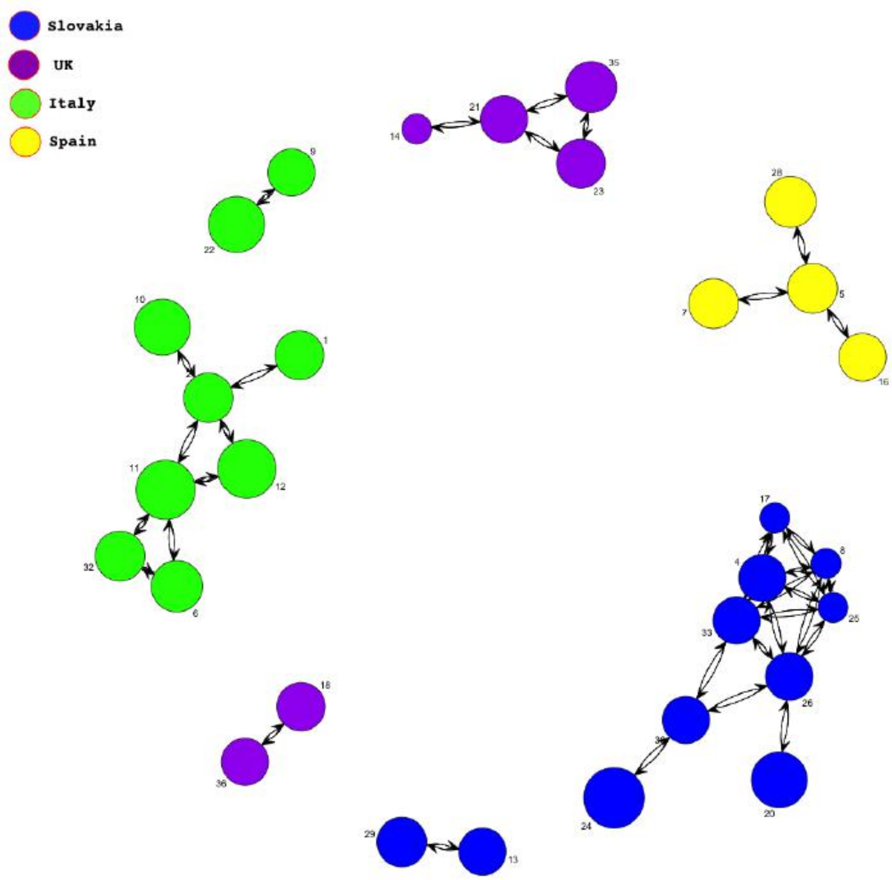


Figure 27: A potential network of local partnerships between the members of the OM community. This network is again essentially a sub-graph of the recommendation network drawn in Figure 24. Here it is assumed that the OM members seek for local collaborators for the same threshold levels required for the cognitive similarity and difference.

References

[AC12] C. C. Aggarwal and Zhai C. *A survey of text clustering algorithms.* Springer, Boston, MA., 2012.

[And12] C. Anderson. *Makers: The New Industrial Revolution.* New York: Crown Business., 2012.

[BA99] AL Barabasi and R Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[DS14] Gregorio D'Agostino and Antonio Scala. *Networks of Networks: The Last Frontier of Complexity.* Understanding Complex Systems. Springer International Publishing, 2014.

[DSZC12] D'Agostino, G., Scala, A., Zlatić, V., and Caldarelli, G. Robustness and assortativity for diffusion-like processes in scale-free networks. *Europhys Lett*, 97(6):68006, 2012.

[Eul36] L. Euler. Solutio problematis ad geometriam situs pertinentis. *Comment. Acad. Sci. U. Petrop.*, 8:128–140, 1736.

[Kle99] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[Koh16] Cindy Kohala. *Making sustainability: How Fab Labs address environmental issues.* PhD Dissertation. Aalto University. Helsinki, Finland., 2016.

[Kon36] Denes Konig. *Theorie der endlichen und unendlichen Graphen.* Lipsia, Akademische Verlagsgesellschaft, 1936.

[LS99] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.

[LS12] D. D. Lee and H. S. Seung. Exploring topic coherence over many models and many topics. *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Association for Computational Linguistics*, pages 952–961, 2012.

[Mal02] Kirsten Malmkjar. *The Linguistics Encyclopedia.* Routledge, 2002.

[New18] Mark Newman. *Networks.* Oxford university press, 2018.

[Pai12] Juliana Paiva. *Towards openness: A study about open design and its translation from theory into practice.* MSc Dissertation, New Media and Digital Culture, University of Amsterdam. Amsterdam, NL., 2012.

[Pow15] A. Powell. *Democratizing production through open source knowledge: from open software to open hardware.* Media, Culture and Society, 2015.

[Ser14] Anna Seravalli. *Making commons: attempts at composing prospects in the opening of production.* PhD Dissertation. New Media, Public Spheres and Forms of Expression at Faculty of Culture and Society, Malmo University. Malmo, Sweden., 2014.

[SS15] Sifatullah Siddiqi and Aditi Sharan. Keyword and keyphrase extraction techniques: A literature review. *International Journal of Computer Applications*, 109(2):18–23, 2015.

[VH05] E. Von Hippel. *Democratizing Innovation.* Cambridge, MA: The MIT Press, 2005.

[WS98] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 1998.