Network of European Union–funded collaborative research and development projects

Michael J. Barber*
Centro de Ciências Matemáticas, Universidade da Madeira, Funchal, Portugal

Andreas Krueger†
Fakultät für Physik, Universität Bielefeld, Bielefeld, Germany

Tyll Krueger‡
Fakultät für Physik, Universität Bielefeld, Bielefeld, Germany and Fachbereich Mathematik,
Technische Universität Berlin, Berlin, Germany

Thomas Roediger-Schluga§
Department of Technology Policy, ARC Systems Research, Vienna, Austria

(Received 20 October 2005; published 29 March 2006)

We describe collaboration networks consisting of research projects funded by the European Union (EU) and the organizations involved in those projects. The networks are substantial in terms of size, complexity, and potential impact on research policies and national economies in the EU. In empirical determinations of the network properties, we observe characteristics similar to those of other collaboration networks, including scale-free degree distributions, small diameter, and high clustering. We present some plausible models for the formation and structure of networks with the observed properties.

DOI: 10.1103/PhysRevE.73.036132 PACS number(s): 89.75.Hc, 89.75.Da

I. INTRODUCTION

Real-world network analysis has recently become a major research topic, following the landmark work of Watts and Strogatz [1]. Most prominent are perhaps the investigations of the structure of the World Wide Web, the network of internet routers, and certain social networks like citation networks. On the theoretical side, one tries to understand the mechanisms of formation of such networks and to derive statistical properties of the networks from the generating rules. On the rigorous mathematical side, there are only a few results for specific models, indicating the difficulty of a purely mathematical approach (for a survey of recent results in this direction, see [2]). Thus, the main approach is to use some mean field assumption to get relevant information about the corresponding graphs. Although it is not clear where the limits of this approach lie, in many cases the results match well with numerical simulations and empirical data. Several useful reviews of recent research in networks are available, such as [3].

In this paper, we study a particular collaboration network. Its vertices are research projects funded by the European Union (EU) and the organizations involved in those projects. In total, the database contains over 20 000 projects and 35 000 participating organizations. The network shows all the main characteristics known from other complex network structures, such as scale-free degree distribution, small diameter, high clustering, and assortative vertex correlations.

Besides the general interest in studying a new, real-world network of large size and high complexity, the study could have a significant economic impact. Improving collaboration between actors involved in innovation processes is a key objective of current science, technology, and innovation policy in industrialized countries. However, little is known about what kind of network structures emerge from such initiatives. Moreover, it is quite likely that network structure affects network functions such as knowledge creation, knowledge diffusion, and the collaboration of particular types of actors. Presumably, this is determined by both endogenous formation mechanisms and exogenous framework conditions. In order to progress in our understanding, it is therefore essential to have sound statistics on the structure of networks we observe and to develop plausible models of how these are formed and evolve over time.

The model networks we use to compare with the empirical data are random intersection graphs, a natural framework for describing projections of bipartite graphs. Discrete intersection graphs similar to the ones we use were first discussed in [4]. We extend and refine the construction from [4] to be more applicable to real-world graphs.

Perhaps the most important finding from our model approach is the strong determination of the real network structure by the degree distribution. That is, most statistical properties we measure in the EU research project networks are the ones observed in a typical realization of a uniform weighted random graph model with given (bipartite) degree distribution as in the EU networks. Since this distribution is characterized by two exponents—one for each partition—we have essentially only four parameters (size, edge number, and exponents) which are needed to describe the entire network. This tremendous reduction of complexity indicates

*Electronic address: mjb@uma.pt
†Electronic address: networks@AndreasKrueger.de
‡Electronic address: tyll.krueger@freenet.de
§Electronic address: Thomas.Roediger@arcs.ac.at

1539-3755/2006/73(3)/036132(13)/$23.00

TABLE I. FP1–FP4 total budget and number of funded projects. The smaller average funding per project and organization in FP4 is an artifact as it involves a large number of scholarships and the like, which are smaller than research projects (however, we cannot isolate the bias created).

Framework Program	Budget[a]	No. of P	Million Euros/P	No. of $(P>1)$[b]	No. of O	Million Euros/O
FP1 (1984–1988)	3.8	3283	1.15	1696	2500	1.52
FP2 (1987–1991)	5.4	3885	1.39	3013	6135	0.88
FP3 (1990–1994)	6.65	5294	1.25	4611	9615	0.69
FP4[c] (1994–1998)	13.3	15061 (9087)	0.88	11374 (8039)	20873	0.64

[a] Billion Euros.

[b] Projects with more then one participating organization.

[c] Research and development projects listed in parentheses. The number excludes all projects devoted to preparatory, demonstration, and training activities.

that only a few basic formation rules are driving the network evolution.

In Sec. II, we describe the preparation of the data on the EU research programs. We present empirical determination of the network properties in Sec. III, followed by an explanation of these properties using a random intersection graph model in Sec. IV. Finally, in Sec. V, we summarize the key results and consider implications of the network properties on EU research programs.

II. THE DATA SET

In this work, we study research collaboration networks that have emerged in the European Union's successive four-year Framework Programs (FPs) on Research and Technological Development. Since their inception in 1984, six FPs have been launched, on the first four of which we have comprehensive data. FPs are organized in priority areas, which include information and communication technologies (ICTs), energy, industrial technologies, life sciences, environment, transportation, and a number of additional activities. In line with economic structural change, the main thematic focus of the FPs has shifted somewhat over time from energy and industrial technologies to the application of ICTs and life sciences. The majority of funding activities are aimed at stimulating research partnerships between firms, universities, research organizations, governmental actors, nongovernmental organizations, lobby groups, etc. Since FP4, the scope of activities has been expanded to also cover training, networking, demonstration, and preparatory activities (for details, see Ref. [5]). In order to keep our data set compatible over the different FPs, we have excluded the latter set of projects from FP4 and focus only on collaborative research projects (see Table I).

In order to receive funding, projects in FP1–FP4 had to comprise at least two organizations from at least two member states. We have retrieved data on these projects from the publicly available Community Research and Development Information Service (CORDIS) projects database [6]. This database contains information on all funded projects as well as a reasonably complete listing of all participating organizations.

The raw data on participating organizations are rather inconsistent. Apart from incoherent spelling in up to four languages per country, organizations are labeled inhomogeneously. Entries may range from large corporate groupings, such as Siemens, or large public research organizations, like the Spanish CSIC, to individual departments or laboratories, and are listed as valid at the time the respective project was carried out. Among heterogeneous organizations, only a subset contains information on the unit actually participating or on geographical location. Information on older entries and the substructure of firms tends to be less complete.

Because of these difficulties, any automatic standardization method akin to the one utilized by Newman [7] is inappropriate to this kind of data. Rather, the raw data have to be cleaned and completed manually, which is an ongoing project at ARC Systems Research. The objective of this work is to produce a data set useful for policy advice by identifying homogeneous, economically meaningful organizational entities. To this end, organizational boundaries are defined by legal control and entries are assigned to the respective organizations. Resulting heterogeneous organizations, such as universities, large research centers, or conglomerate firms are broken down into subentities that operate in fairly coherent areas of activity, such as faculties, institutes, divisions, or subsidiaries. These can be identified for a large number of entries, based on the available contact information of participants, and are comparable across organizations.

The case of the French Centre National de la Recherche Scientifique (CNRS), the most active participant in the EU FPs, may serve as an illustration. First, 785 separate entries were summarized under a unique organizational label. Next, these 785 entries were broken down into the eight areas of research activity in which CNRS is currently organized. Based on available information on participating units and geographical location, 732 of the 785 entries could be assigned to one of these subentities. For the remaining 53 entries, the nonspecific label CNRS was used.

Comparable success rates were achieved for other large public research organizations and universities. Due to scarcer information, firms could not be broken down at a comparable rate. Moreover, due to resource constraints, standardization work has focused on the major players in the FPs. Organizations participating in fewer than a total of 30 projects in

TABLE II. Basic network properties of FP1–FP4 organizations projection.

Graph characteristic	FP1	FP2	FP3	FP4
No. of vertices N	2500	6135	9615	20873
(N for largest component)	(2038)	(5875)	(8920)	(20130)
N outside largest component	462	260	695	743
No. of edges M	9557	64300	113693	199965
(No. of edges M largest component)	(9410)	(64162)	(113219)	(199182)
Mean degree $\bar{d}$	7.65	20.96	23.65	19.16
($\bar{d}$ largest component)	(9.23)	(21.84)	(25.39)	(19.79)
Maximal degree $d_{\max}$	140	386	648	649
Mean triangles per vertex $\triangle$	22.90	169.70	244.91	146.04
($\triangle$ largest component)	(27.97)	177.16	263.84	151.26
Maximal triangle number	966	5295	15128	10730
Cluster coefficient $\bar{C}$	0.57	0.72	0.72	0.79
($\bar{C}$ largest component)	(0.67)	(0.74)	(0.75)	(0.81)
Number of components	369	183	455	467
Diameter of largest component	9	7	9	10
Mean path length λ of largest component	3.70	3.27	3.32	3.59
Exponent of degree distribution	−2.1	−2.0	−2.0	−2.1
Variance of degree exponent	0.4	0.3	0.3	0.3
Exponent of organization size distribution	−2.1	−1.9	−1.7	−1.8
Variance of size exponent	0.5	0.3	0.5	0.3
Mean no. of projects per organization $\mathbb{E}(\|O\|)$	2.40	4. 87	5.6	6.24
Maximal size ($\max\|O\|$)	130	82	138	172

FP1–FP4 have not been broken down yet. Due to these limitations in processing the data, we cannot rule out the possibility of a bias in analyzing our data. However, we have run all the reported analyses with the undivided organizations and have obtained qualitatively similar results, apart from different extreme values, e.g., maximum degree.

Table I displays information on the present data set, which contains information on a total of 27 758 projects, carried out over the period 1984–2004. It shows that the total budget as well as number of funded projects has increased dramatically from FP1 to FP4. Moreover, it provides a rough measure on the completeness of the available data. For a sizable number of projects, the CORDIS project database lists information only on the project coordinator. This is due to the age of the data and inhomogeneous disclosure policies of different units at the European Commission. Comparing the number of projects containing information on more than one participant with the total number of projects funded in each FP shows that the data are fairly complete as of FP2.

The facts that FP1 was the first program launched and that the available data are rather incomplete make it exceptional in many respects. We therefore focus our analyses on FP2–FP4 and only give graph characteristic values for FP1 to indicate the difference from the networks created by the subsequent FPs.

III. THE NETWORK STRUCTURE

In this section, we present the basic properties of the network structure for projects and organizations in the first four EU Framework Programs. We consider both graphs as intersection graphs [4], each being the dual of the other, which, for our purposes, is generally more convenient than the usual bipartite-graph point of view. The vertices of an intersection graph are given by an enumerated collection of sets with elements from a given fixed base-set, while the edges are defined via an intersection property (edge $\triangleq$ nonempty intersection of two sets). The sets need not be distinct.

We denote by $\mathcal{P}=\{P_1;\ldots;P_M\}$ the family of projects and by $\mathcal{O}=\{O_1;\ldots;O_N\}$ the family of organizations. Projects are understood as labeled sets of organizations and organizations as labeled sets of projects. The corresponding intersection graphs are denoted by G_P and G_O; we will also use the terms P graph and O graph for them. The size $|x|$ of a vertex x from G_P or G_O is the cardinality of the set corresponding to the vertex; in the picture of bipartite graphs, the size is just the degree of the vertex. In Tables II and III, we give some basic parameters measured on the P and O graphs from the four Framework Programs. Since the degree distribution for P graphs is a superposition of two power-law distributions (one for small degree values and one for large values), we give the corresponding values for the exponents parenthetically. The clustering coefficients shown are defined (following [1]) as follows. Assume that vertex v has d_v neighbors; potentially, $d_v(d_v-1)/2$ edges could exist between those neighbors, forming triangles. Define an auxiliary, vertex-specific clustering coefficient C_v as the ratio of the number of those triangles actually formed to the number of triangles that potentially could be formed. The clustering coefficient for the

TABLE III. Basic network properties of FP1–FP4 projects projection.

Graph characteristic	FP1	FP2	FP3	FP4
No of. vertices N	3283	3884	5528	9087
(N for largest component)	(2764)	(3662)	(5027)	(8566)
N outside largest component	519	222	501	521
No. of edges M	51217	94527	202358	348542
(No of. edges M largest component)	(50940)	(94471)	(202306)	(348474)
Mean degree $\bar{d}$	31.20	48.68	73.20	76.71
($\bar{d}$ largest component)	(36.86)	(51.60)	(80.49)	(81.36)
Maximal degree $d_{\max}$	282	387	917	771
Mean triangles per vertex $\triangle$	774.41	871.19	1970.30	2034.31
($\triangle$ largest component)	919.53	923.98	2167.05	2158.03
Maximal triangle number	12903	11125	37247	41141
Cluster coefficient $\bar{C}$	0.67	0.54	0.44	0.47
($\bar{C}$ largest component)	(0.75)	(0.57)	(0.48)	(0.50)
Number of components	369	183	455	467
Diameter of largest component	9	7	10	9
Mean path length λ of largest component	3.24	2.80	2.72	2.80
Exponent of degree distribution	(−0.8,−3.4)	(−0.7,−3.3)	(−0.6,−3.7)	(−0.3,−2.2)
Variance of degree exponent	(0.4, 3.6)	(0.3, 1.7)	(0.3, 1.4)	(0.2, 0.6)
Exponent of project size distrbution	−3.59	−2.9	−3.4	−4.1
Variance of size exponent	0.6	0.4	0.2	0.3
Mean no. of organizations per project $\mathbb{E}(\|P\|)$	3.15	3.08	3.22	2.71
Maximal size ($\max\|P\|$)	20	44	73	54

network as a whole is just the average of C_v over all vertices in the graph.

As expected, FP1–FP4 are of small-world type: high clustering coefficient and small diameter of the giant component. There is a slight increase in the clustering coefficient of the O graphs from FP1 to FP4, indicating a stronger integration among groups of collaborating organizations. This is also reflected in the mean organization size which increases from 2.4 to 6.2. There is an interesting jump in the P graph mean degree values and the mean triangle numbers between FP1 and FP2 and between FP2 and FP3. The maximal degrees of the O graphs are high in comparison with the mean degrees, which is a consequence of the power-law degree structure. For the P graphs, the gap between mean and maximal degree is less pronounced.

More information is contained in the statistical properties of the relevant distributions. The numerical data strongly indicate that the size distributions follow power laws. Also, the O graph degree distribution is of power-law type, while the project-graph degree distribution is a superposition of two scale-free distributions, one dominating the distribution for small degree values (up to 100) and one relevant for the large degree values. We discuss these properties at greater length in the following sections.

A. Size distributions

The size distributions are the basic distributions for the EU networks since, as will be shown in Sec. IV B, a typical sample from the random graph space with fixed size distributions as in FP2–FP4 will have similar statistical properties to FP2–FP4. This strongly suggests that there is essentially no additional correlation in the data once the size distribution is known. Both the O graph and P graph size distributions show clear asymptotic power-law distributions for FP1–FP4 (Figs. 1 and 2). In terms of the corresponding bipartite graph, these are just the degree distributions of the project and or-

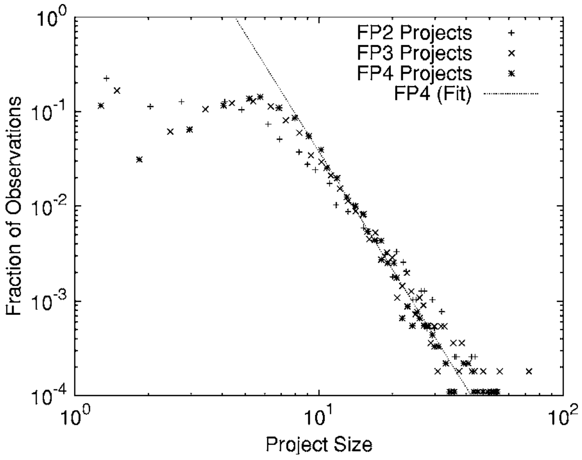

FIG. 1. Distribution of project sizes. The size of a project is defined as the number of organizations taking part in the project. The tails of the distributions are power laws; for FP4, we show a power-law fit to the data with exponent −4.1.

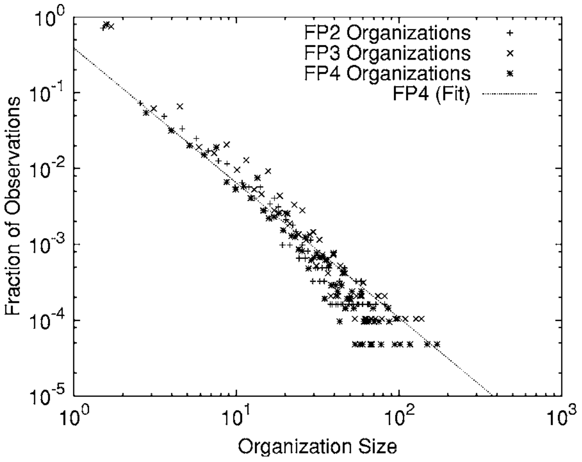

FIG. 2. Distribution of organization sizes. The size of an organization is defined here as the number of projects in which it takes part. The tails of the distributions are power laws; for FP4, we show a power-law fit to the data with exponent −1.8.

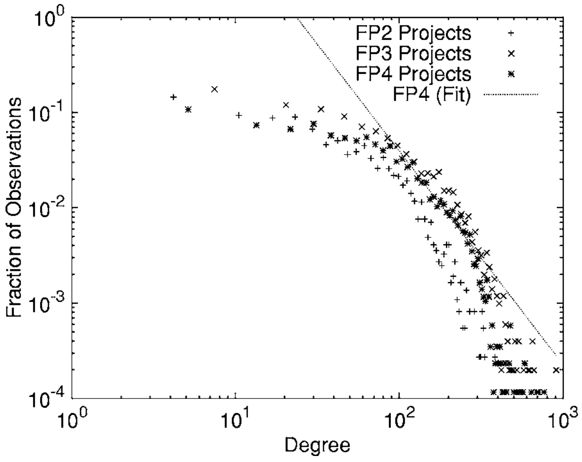

FIG. 3. Degree distribution of projects projection. The distributiond show a structure formed from the superposition of two power laws; for FP4, we show a power-law fit to the high-degree data with exponent −2.2.

ganization partitions. While the O graph size distribution is of power-law type over the whole size range, the P graph size distribution deviates strongly from the power law for small size values. In Sec. IV, we give a possible explanation for the appearance of the power-law distribution for size.

The numerical values for the exponents of the organization size distributions from FP2 to FP4 are slightly below 2, but constant within the error tolerance. This indicates that the distribution of organizations able to carry out a particular number of projects has not changed in the three Framework Programs. A complementary interpretation of this finding is that the underlying research activities, which we know to have changed over time, have not altered the mix of organizations participating in a particular number of projects in each Framework Program. It is further worth noting that the values of the O graph exponents are close to the critical value 2; hence the size expectation could diverge for large graphs (whether the value is really below 2 or not is still unclear due to the error tolerance).

The picture is similar for the P graphs, although there are some differences in the initial behavior (that is, for small project sizes) and in the exponent value. The value of the local minima at size 2 decreases from FP2 to FP4. This points to the existence of an optimal project size within the regime of the EU FPs. Moreover, the rise in the average project size indicates that increases in the available funding from FP2 to FP4 lead to not only more projects, but also slightly larger projects. This is consistent with recommendations from evaluation studies and the stated attempts of the EU commission to reduce its administrative burden. As a whole, the size distribution for the P graphs in the asymptotic regime matches well to a power law with exponent around −3, hence indicating that the mechanisms for coagulation of organizations into a project did not greatly change from FP2 to FP4.

B. The degree distribution

Since the degree distribution in the projection graphs is just the distribution of the sizes of the 2-neighborhoods consisting of the sets of next-nearest neighbors in the bipartite graph, it is not surprising that this quantity is closely connected to the size distribution. In the absence of other special correlations, it can be shown (see Sec. IV) that the degree distribution is determined by the size distribution in a rather simple way; namely, for the case when both size distributions are scale-free with exponents, say α (O size) and β (P size), the P graph degree distribution is a superposition of two power-law distributions with exponents $\alpha-1$ (and cutoff given by the maximal O-size value) and β. An analogous property holds for the O graph.

In Figs. 3 and 4, we show the degree distributions for the P and O graphs in a log-log plot. While the organization graphs for FP2–FP4 show a clear power law, the picture for the project graphs is more complicated. As previously mentioned, the P graph degree distribution shows two different power laws, one for the initial segment up to degree 150 and another one for large degrees. Nevertheless, there is still a widely scattered heavy tail in the degree distribution.

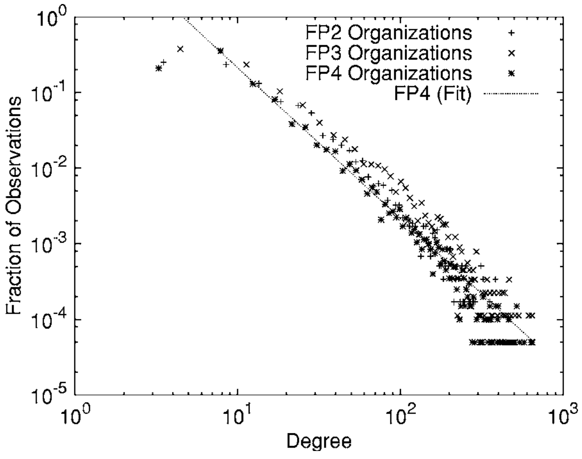

FIG. 4. Degree distribution of organizations projection. The tails of the distributions are power laws; for FP4, we show a power-law fit to the data with exponent −2.0.

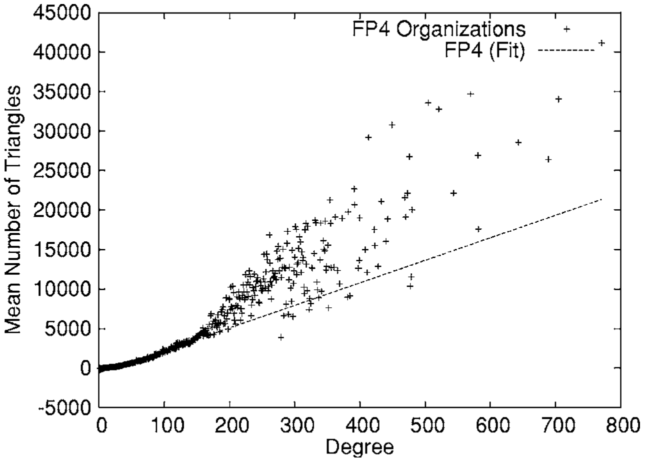

FIG. 5. Relation between degree and number of triangles in the projects projection. For each degree value, we show the mean number of triangles, conditioned on the vertices with the given degree. For low degree values, a strong linear relationship is observed, but the strength of the relationship weakens with high degrees. Here, we show only the data for FP4 for comprehensibility; similar results hold for the other Framework Programs.

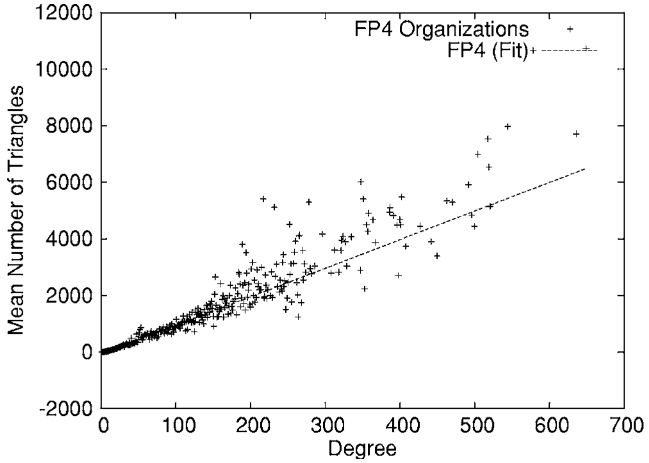

FIG. 6. Relation between degree and number of triangles in the organizations projection. For each degree value, we show the mean number of triangles, conditioned on the vertices with the given degree. For low degree values, a strong linear relationship is observed, weakening slightly with higher degrees (compare with Fig. 5). Here, we show only the data for FP4 for comprehensibility; similar results hold for the other Framework Programs.

C. Clustering, correlation, and edge multiplicity

By their construction process, intersection graphs have a naturally high clustering coefficient, since an organization which participates in, say, k projects generates a complete subgraph of order k in the P graph among these projects. If the probability for an organization to be in more than one project is asymptotically bound away from zero, it follows that the P graph (and similarly for the O graph through an analogous argument) has a nonvanishing clustering coefficient. In the present study, we focus on the triangle number $\triangle(x)$, defined as the number of triangles in the ($\mathcal{P}$ or $\mathcal{O}$) graph containing x, as a measure of local clustering. We define the degree-conditional mean triangle number as $\triangle_k := \mathbb{E}\{\triangle(x)|d(x)=k\}$, where $d(x)$ is the degree of vertex x. As seen in Figs. 5 and 6, we have $\triangle_k \sim k$ for both graph types.

There is a good explanation for this type of behavior in the framework of intersection graphs (see Sec. IV). As noted above, high clustering in intersection graphs is not necessarily an indication of local correlations between vertices. This is already seen in the case of an Erdös-Renyi random bipartite graph where an edge between any project and organization is drawn in an independent, identically distributed (i.i.d.) fashion with probability p. If $\mathcal{P}$ and $\mathcal{O}$ are of equal cardinality N and $p=\frac{c}{N}$, the expected bipartite degree equals c. For large N a typical realization of the random graph looks locally like a tree with branching number $c-1$. However, for the projection graphs, we obtain a positive clustering coefficient that is independent of N, since most projects and organizations cause complete graphs of order c and a typical vertex is therefore a member of order c cliques, each of order c.

A better indication for the presence of correlations is given by the so-called multiplicity of edges. For a link between two organizations or projects it is sufficient to have just one project or organization, respectively, in common, but of course there could be more. Given an edge $x \sim y$, we define $m(x,y) := |x \cap y| - 1$ and call it the multiplicity of the edge. As will be discussed in the next section, random intersection graphs without local search rules can nevertheless admit a high edge multiplicity. In Figs. 7 and 8, the multiplicity distribution is shown for P and O graphs of FP2–FP4. There is an almost perfect power-law behavior with exponent 4.3. Note that positive multiplicity in the projection graphs translates in the bipartite graph picture into the presence of cycles of length 4. The presence of exceptionally high multiplicity in the P graphs may be caused by memory effects due to prior collaborative experience. Also, a greater edge

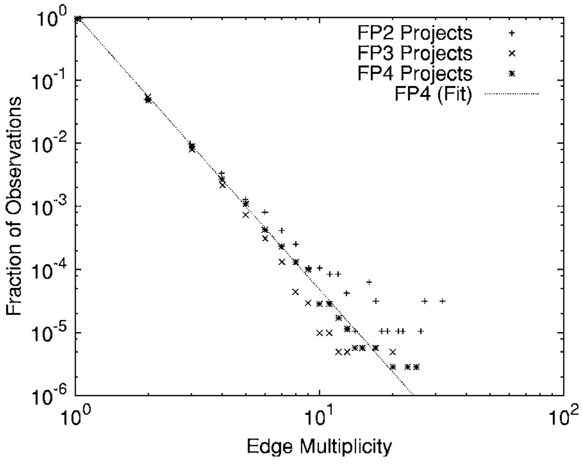

FIG. 7. Distribution of edge multiplicities in the projects projection. An almost perfect power-law distribution is observed for all Framework Programs. The multiplicity is strongly indicative of correlations in the edge formation rules, possibly caused by memory effects due to prior collaborative experiences amongst the participating organizations or by the fact that organizations are active in a wider set of complementary activities.

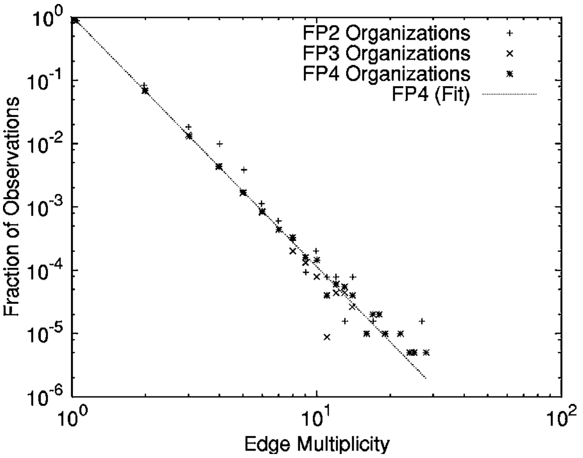

FIG. 8. Distribution of edge multiplicities in the projects projection. An almost perfect power-law distribution is observed for all Framework Programs. As with the P graphs in Fig. 7, the high multiplicities are indicative of correlations in the edge formation rules.

multiplicity may result from the fact that organizations are active in a wider set of complementary activities. In this case, intraorganizational links and knowledge flows may also be of importance, as the search for potential partners may be influenced by the collaboration behavior of other actors within an organization. Such effects should be detectable from a fine structure analysis of the time evolution of the corresponding graphs.

D. Diameter and mean path length

There is essentially no difference in the diameter value of the largest component in the four Framework Program networks. A classical random graph of the same size and the same edge number would have a diameter about $log_{\bar{d}}N$, where N is the number of vertices and $\bar{d}$ is the average degree of the vertices. The mean path length is about one-third of the diameter and shows a slightly higher variation between the different framework programs. It is well known that the expected path length in random graphs with a scale-free degree distribution and exponent less than 3 is essentially independent of the graph size (the diameter of the largest component still increases in N but only as log log$N$). The same holds for random intersection graphs with power-law size and degree distributions. Since the O graphs seem to fall into that class, the almost constant diameter and path length is not surprising. Although the P graphs do not show an asymptotic power-law structure for the degree, there is a strong increase in the edge density from FP2 to FP4, keeping the diameter of the largest component almost fixed.

IV. A RANDOM INTERSECTION GRAPH MODEL

Intersection graphs are a natural framework for networks derived from a membership relation, such as citation networks, actors networks, or networks reflecting any other kind of cooperation. As previously mentioned, intersection graphs by construction have a high clustering coefficient. As explained below, the clique distribution of a random intersection graph is almost given by the size distribution of the dual graph.

A. Random intersection graphs with given size distribution

One of the simplest random intersection models is constructed in the following way. Knowing the size of a set to be constructed, we generate a random subset from a finite base set $X=\{a_1,a_2,\ldots,a_N\}$ of N elements, such that each set element is drawn i.i.d. uniformly from X. These subsets constitute the vertices of a random graph. Edges are defined via the set intersection property, namely, we have an edge between i and j (denoted by $i\sim j$) if and only if the associated subsets A_i and A_j have nonempty intersection (to compare with earlier sections, A stands here for either projects sets P or organization sets O). The size (cardinality) of the subsets is either itself a random variable drawn i.i.d. from a probability distribution $\varphi(k)$ or given by a list $D_k := |\{A_i : |A_i| = k\}|$ (where for each i a conditional random choice is made to which size class it belongs). For the latter case, we define again $\varphi(k) := \frac{D_k}{M}$ where M is the total number of sets to be formed.

Since we want to compare the model with the EU collaboration networks, we are mainly interested in the situation when φ is an asymptotic power-law distribution

$$\varphi(k) = \frac{1}{k^{\alpha+o(1)}}, \quad \alpha > 2. \tag{1}$$

This assumption is also reasonable for many other applications where vertices are formed from a base set of elements. To obtain an interesting limiting random graph space, we further assume that the number of chosen subsets is C_1N where C_1 is neither too large nor too small (for FP2–FP4 we have about twice as many organization as projects, hence C_1 is either 2 or 0.5).

A basic quantity for the analysis of intersection graphs is $P_{k,l}(N)$, the conditional edge probability given the size of two subsets:

$$P_{k,l}(N) := \Pr\{i \sim j \| A_i| = k \text{ and } |A_j| = l\} \tag{2}$$

$$= \Pr\{A_i \cap A_j \neq \varnothing \| A_i| = k \text{ and } |A_j| = l\} \tag{3}$$

$$= 1 - \frac{\binom{N-k}{l}}{\binom{N}{l}} \tag{4}$$

$$= 1 - \frac{(N-k)!(N-l)!}{N!(N-k-l)!} \tag{5}$$

$$= 1 - \frac{(N-k)(N-k-1)\cdots(N-k-l+1)}{N(N-1)(N-2)\cdots(N-l+1)}. \tag{6}$$

Using the condition $lk \ll N$, we obtain

$$P_{k,l}(N) = 1 - \frac{\left(1-\frac{k}{N}\right)\left(1-\frac{k+1}{N}\right)\cdots\left(1-\frac{k+l-1}{N}\right)}{\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{l-1}{N}\right)} \tag{7}$$

$$= 1 - \frac{1 - \frac{lk+\frac{1}{2}l(l-1)}{N} + o\left(\frac{1}{N}\right)}{1 - \frac{l(l-1)}{2N} + o\left(\frac{1}{N}\right)} \tag{8}$$

$$= \frac{lk}{N} + o\left(\frac{1}{N}\right). \tag{9}$$

With this result, we can easily calculate the conditional degree distribution for a vertex of given size. First, we estimate the conditional subdegree distribution $\psi_l(k,m)$ with respect to a given group of vertices of size m. Here, the subdegree $d_m(i)$ of a vertex i is defined as the number of edges i has with vertices of size m. Clearly, the subdegrees are related to the degree $d(i)$ through $d(i)=\Sigma_m d_m(i)$. We have

$$\psi_l(k,m) \coloneqq \Pr\{d_m(i)=k \| A_i| = l\} \tag{10}$$

$$= \sum_G \Pr\{|\{j\|A_j| = m\}| = G\}\binom{G}{k} \times \left[\frac{ml}{N} + o\left(\frac{1}{N}\right)\right]^k \left[1 - \frac{ml}{N} + o\left(\frac{1}{N}\right)\right]^{G-k}. \tag{11}$$

The probability that a randomly chosen vertex j has size m equals, by assumption, $C_2/m^{\alpha+o(1)}$ with the normalization constant C_2 defined by $1=\Sigma_m C_2/m^{\alpha+o(1)}$. We therefore obtain

$$\psi_l(k,m) = \lim_{N\to\infty}\binom{C_1 N \frac{C_2}{m^\alpha}}{k}\left[\frac{ml}{N} + o\left(\frac{1}{N}\right)\right]^k \times \left[1 - \frac{ml}{N} + o\left(\frac{1}{N}\right)\right]^{C_1 N C_2/m^\alpha - k}, \tag{12}$$

which converges to a Poisson distribution

$$\psi_l(k,m) = \frac{c(m)^k}{k!}e^{-c(m)} \tag{13}$$

with $c(m)=m^{1-\alpha}lC_1C_2$. Since the distribution $\psi_l(k)$ of the degree of vertices i with $|A_i|=l$ is the convolution of the Poisson distributions $\psi_l(k,m)$, we obtain again a Poisson distribution for $\psi_l(k)$:

$$\psi_l(k) = \frac{c_l^k}{k!}e^{-c_l} \tag{14}$$

with $c_l=\Sigma_m c(m)=lC_3$, where $C_3=\Sigma_m m^{1-\alpha}C_1C_2$ is a well-defined constant since $\alpha>2$.

The total degree distribution $\psi(k)$ remains to be estimated. In [8], conditions were given describing when a superposition of Poisson distributions results in a scale-free distribution. Specifically, we get the following asymptotic estimate:

$$\psi(k) = \sum_m \varphi(m)\frac{(mC_3)^k}{k!}e^{-mC_3} \tag{15}$$

$$= \sum_m \frac{1}{m^{\alpha+o(1)}}\frac{(mC_3)^k}{k!}e^{-mC_3}. \tag{16}$$

The main contribution to $\psi(k)$ comes from a rather small interval of m values, called $I_{ess}(k)$. This interval has the property that for $m \in I_{ess}(k)$, the expectation $\mathbb{E}[d(i)\rangle|A_i|=m]$ is of order k. The exponential decay of the Poisson distribution guarantees that the remaining parts of the sum become arbitrarily small for large k. It is important that the constant c_l has a linear l dependence since an l proportionality with exponent larger than 1 would force the degree distribution to have gaps due to a lack of overlap of the individual Poisson distributions. We therefore obtain for the degree distribution a power law with the same exponent α as in the size distribution.

Although the intersection model gives a power-law degree distribution when the size distribution is already of power-law type, we will not obtain a power-law distribution for the size on the dual graph unless additional assumptions are made on the set formation rules. It is easy to see that the size distribution on the dual graph is asymptotically Poisson. Since $\Pr\{|x|=k\} \sim \binom{M}{k}\left(\frac{\mathbb{E}(|A|)}{N}\right)^k\left(1-\frac{\mathbb{E}(|A|)}{N}\right)^{M-k}$ and $\mathbb{E}(|A|)$ converges as well as $\frac{M}{N}$ for $M,N\to\infty$, we obtain in the limit a Poisson distribution. Nevertheless, the degree distribution on the dual graph still admits a scale-free part induced by the scale-free size distribution of the intersection graph. We will not discuss many of the details, but instead provide a simple estimation for the lower bound on the number of elements a_i with $d(a_i)=k$. Namely, the number of elements a_i which are members of sets A_j with $|A_j|=k$ is for large k and $M,N \gg k$ about $\frac{kM\times\text{const}}{k^\alpha}=\frac{N\times\text{const}}{k^{\alpha-1}}$. Since $d(a_i)\geqslant k$ for $a_i \in A_j$ with $|A_j|=k$, we obtain $\frac{\text{const}}{k^{\alpha-2}}$ as a lower bound on the density of elements a_i with degree greater than or equal to k (note that we assumed $\alpha>2$). This estimate holds of course only up to the maximal size value k, which is in the range of the power law distribution for the set sizes $|A_i|$. For larger k values there is a rapid exponential decay.

The last argument clarifies also the situation when one wants to impose conditions on the size distribution and the dual size distribution. Without going into the details of the rather involved analysis, we simply state that the resulting degree distribution is given by a superposition of the size distibution and the dual size distribution (the last one enters with an exponent reduced by 1). This explains essentially the picture for the degree distribution for the P graph.

Finally, we consider the mean triangle (conditioned on the degree) degree dependence, which shows a clear linear behavior in the empirical data. We argue that this is again a consequence of the power-law distribution for the size. First observe that a size k element $a_i \in A_j$ induces a $k-1$ complete subgraph on the neighborhood vertices of A_j. Furthermore, each maximal k clique in which A_j is a member generates $(k-1)(k-2)/2$ triangles for A_j. Since the size distribution of the elements a_i is Poisson with expectation of, say, c and the degree of A_j is proportional to the size $|A_j|$, we obtain for the conditional expected number of triangles $\triangle_k$ given the degree k:

$$\triangle_k := \mathbb{E}(\text{number of triangles containing } A|d(A)=k) \sim \frac{c^2}{2}\text{const} \times k. \tag{17}$$

In deriving Eq. (17), we used the facts that with high probability the size of the intersection between two sets A_i and A_j has cardinality 1 (conditioned on the two sets having a nonempty intersection) and that the Poisson distribution has an exponentially decaying tail.

B. A Molloy-Reed version of random intersection graphs and a Bernoulli-type model

We sketch the construction of random intersection graphs with given size distribution φ and size distribution ψ on the dual. The two distributions are not independent but must satisfy the condition $\Sigma_i i\varphi(i) = \Sigma_i i\psi(i)$. There are further restrictions on the maximal size in order to get a reasonable random graph model. Note that the problem is equivalent to the construction of a random bipartite graph given the degree sequence on the two partitions. The approach we follow is a variation of the graph construction algorithm usually attributed to Molloy and Reed [9] (actually given earlier by Bollobás [10]).

Assign first to each set A and each element a from the base set a random size value according to the given distributions φ and ψ. Let D_k be the resulting set of elements a_i with size k. Replace each element from D_k by k virtual elements $a_{i,l}, l=1,2,\ldots,k$ and form a new base set X' with all the virtual elements. The set formation process for the sets $\{A_i\}$ is now the same as in the previous section except that each chosen virtual element $a_{i,l}$ will be removed from X' when it was selected first into a set. After the sets are constructed we identify the virtual elements back into the original ones, remove multiple and self-links, and define the corresponding set graph in the usual way.

By construction the resulting size distribution on the dual graph will be given by ψ as long as the probability of choosing two virtual elements $a_{i,l}$ and $a_{i,m}$ (corresponding to the same element a_i) is sufficiently small. To ensure this one has to impose restrictions on the maximal size values. It is not difficult to show that the correlation between the size of A and the size of an element a is multiplicative. In case of a linear relation between the number of sets N and the number of elements M we have

$$\Pr\{a \in A||A| = k \text{ and } |a| = l\} \sim \frac{\text{const}}{N}kl. \tag{18}$$

To see this observe that

$$\Pr\{a \in A||A| = k \text{ and } |a| = l\} = 1 - \Pr\{\text{among the } k \text{ choices to generate } A \text{ is no virtual } a \text{ element}\} \tag{19}$$

$$= 1 - \frac{M^* - l}{M^*}\frac{M^* - 1 - l}{M^* - 1}\cdots\frac{M^* - k - l + 1}{M^* - k + 1} \tag{20}$$

with M^* being the number of virtual elements. The last formula has the same structure as the expression for the pairing probability in the previous section, hence we get, for $lk \ll M^*$ and bounded first moments of the ψ distribution, the claimed multiplicative correlation. We note that there is also a variant of the Molloy-Reed construction which produces an additive size-size correlation such that $\Pr\{a \in A||A|=k \text{ and } |a|=l\} \sim \frac{\text{const}}{N}(k+l)$ holds (see [11] for details of the algorithm).

We next present a simulation-based comparison of the multiplicative and additive Molley-Reed model with the FP4 network. The input size distributions for the Molloy-Reed simulations are the same as in FP4. For completeness we also include the simulation results based on the simple random intersection graph model defined in the previous section. To make clear which size distribution is given in that case we use the notation P model (O model) for the intersection graph with fixed P (O) size distribution and denote by PO model the corresponding Molloy-Reed graphs since both size distributions are fixed therein. Figures 9 and 10 show the degree distribution for the O and P graphs. There is excellent agreement between the real FP4 network projections and typical samples of the multiplicative Molloy-Reed model over the whole range of degree values. This is quite remarkable since a considerable bias from the almost independence of the Molloy-Reed model should be visible in the degree distributions. The fact that there is no deviation between the degree distributions indicates that the majority of project-organization alignments is essentially a random process. Furthermore, the additive model reproduces the FP4 P graph degree distribution only well for large degree values indicating that the correlation is indeed multiplicative.

Two quantities measuring local correlations are the triangle degree dependence and the distribution of edge multiplicity introduced earlier. Figure 11 compares the triangle degree correlation for the O graph. Although the overall picture is similar (linear dependence up to medium degree) there is a clear tendency for higher triangle numbers in FP4 for large degree values. Again the multiplicative version matches better with the data than does the additive model.

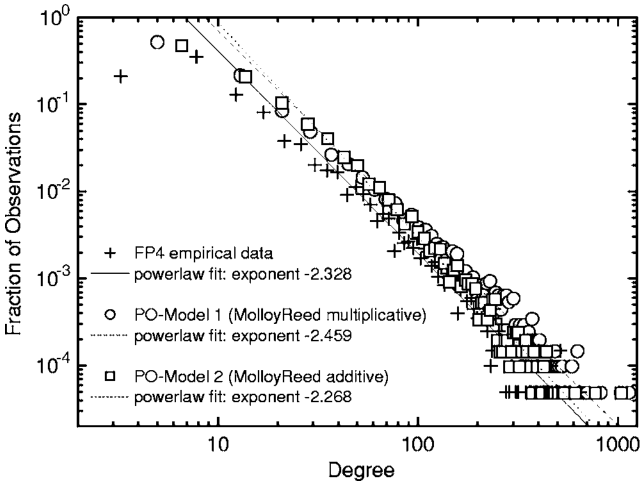

FIG. 9. Simulated degree distribution for the O graphs. The empirical FP4 data are the same as in Fig. 3. The PO model takes as input the empirical organization sizes and project sizes, and randomly pairs an organization to a project using the Molloy-Reed algorithm described in Sec. IV B. During that pairing, both the multiplicative and the additive degree-degree correlations produce networks that are very similar to the empirical O graph with respect to the degree.

The edge multiplicity—again for the O graphs—is shown in Fig. 12. The real graph has a considerably smaller value in the exponent and extends to almost twice as large a maximal multiplicity value. Nevertheless, both Molloy-Reed models show a sharp scale-free distribution for the multiplicity. This is quite surprising, since, naively, one would expect the probability for positive edge multiplicity to go to zero as N becomes large. In summary, one has a strong agreement between the real data and the multiplicative Molloy-Reed model (the comparison results for FP2 and FP3 are almost identical to the situation with FP4 and have therefore not been depicted here). Only in the fine structure of clustering characteristics are some differences observed.

Finally, we briefly outline why, under certain circumstances, almost independent models like the Molloy-Reed one can have a scale-free edge multiplicity distribution. To keep the discussion as transparent as possible, we study the question in a pure bipartite Bernoulli model, which can be thought of as a kind of predecessor to the Cameo model discussed below.

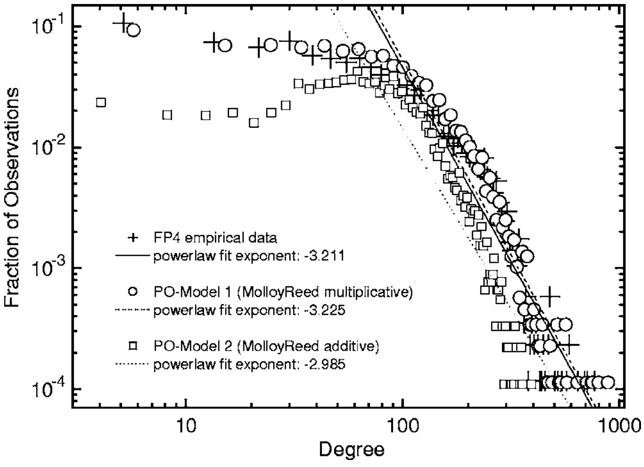

FIG. 10. Simulated degree distribution for the P graphs. The empirical FP4 data are the same as in Fig. 4. With respect to the degree, the Molloy-Reed algorithm with multiplicative degree-degree correlation produces a network that more closely matches the empirical P graph than the modified Molloy-Reed algorithm with additive degree coupling.

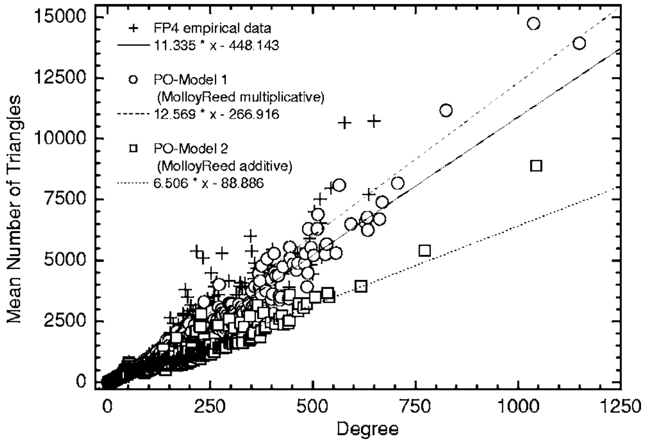

FIG. 11. Simulated triangle degree dependence for the O graphs. The empirical FP4 data are the same as in Fig. 6. With respect to the mean triangle number (conditioned on the degree), the Molloy-Reed algorithm with multiplicative degree-degree correlation produces a network more similiar to the empirical O graph than the modified Molloy-Reed algorithm with additive degree coupling.

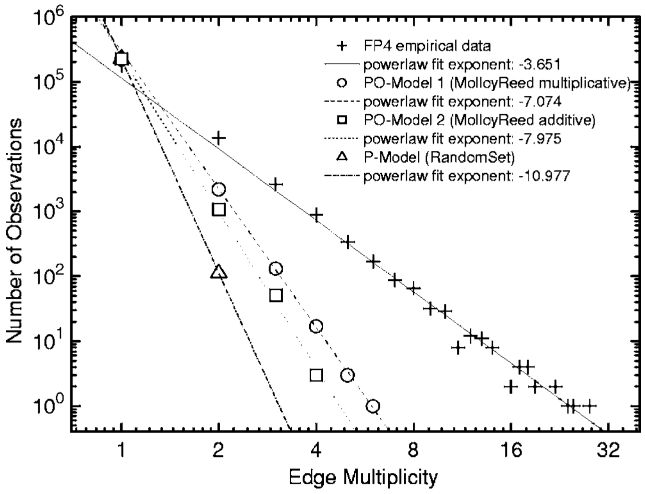

FIG. 12. Simulated edge multiplicity for the O graphs. The empirical FP4 data are the same as in Fig. 8. The two Molloy-Reed algorithms are unable to generate networks reproducing the empirically observed edge multiplicity, in terms of either the exponent or the absolute numbers; the empirical case has nonrandom features that a more advanced model needs to imitate. Also shown is a P model network, in which only the empirical project sizes are taken as input for random sets of organizations; the organization sizes automatically form a Poisson-like distribution. The P model has even smaller edge multiplicities. All models show scale-free edge multiplicities.

To each vertex from the O and P partitions (with cardinality N and M), we assign a power-law distributed, positive integer parameter $\mu(P)$ and $\nu(O)$ with exponents α and β. That is we partition the P and O vertices into sets $D_\mu := |\{P|\mu(P)=\mu\}|$ and $G_\nu := |\{O|\nu(O)=\nu\}|$ such that $|D_\mu| = \frac{C_P M}{\mu^\alpha}$ and $|G_\nu| = \frac{C_O N}{\nu^\beta}$ where C_P and C_O are normalization constants. We further assume that M and N are proportional with $C_{op}=M/N$, and put

$$\Pr\{P \sim O\} := \frac{c}{N}\mu(P)\nu(O). \tag{21}$$

In Eq. (21), c is a free parameter; the ratio c/N regulates the number of edges realized in the network. It is easily shown that the expected degree, conditioned on the μ or ν value, is proportional to μ or ν, respectively, and therefore the (bipartite) degree distribution on each partition has the same exponent as μ or ν. Note that the maximal μ and ν values are given by $\mu_{\max} \sim M^{1/\alpha}$ and $\nu_{\max} \sim N^{1/\beta}$.

Since the edge multiplicity in the projection graph corresponds to the number of paths of length 2 in the bipartite graph, we define $E_k^{(P2)} := \mathbb{E}|\{(P,P')$: there are exactly k paths of length 2 between P and $P'\}|$ and $E^{(P2)} := \sum k E_k^{(P)}$. For fixed P and P' with parameters μ and μ' the expected number of paths of length 2 between the two vertices is given by

$$\sum_\nu \frac{c^2}{N^2}\mu\mu'\nu^2|G_\nu| \tag{22}$$

and therefore the expected total number of 2-paths in the P partition is

$$E^{(P2)} = \sum_{\mu,\mu'} |D_\mu||D_{\mu'}| \sum_\nu \frac{c^2}{N^2}\mu\mu'\nu^2|G_\nu| \tag{23}$$

$$= \sum_{\mu,\mu'} \sum_\nu \frac{C_O C_P^2 M}{C_{op}(\mu\mu')^{\alpha-1}\nu^{\beta-2}}. \tag{24}$$

On the other hand, we have for the probability of an edge between P and P' in the P-projection graph the estimate

$$\Pr\{P \sim P'\} = 1 - \prod_\nu \left(1 - \frac{c^2}{N^2}\mu\mu'\nu^2\right)^{|G_\nu|} \tag{25}$$

$$\simeq 1 - \exp\left(-\sum_\nu \frac{C_O c^2 \mu\mu'}{C_{op} M \nu^{\beta-2}}\right) \tag{26}$$

and hence for the expected total number of edges E

$$E \simeq \sum_{\mu,\mu'} \frac{C_P^2 M^2}{(\mu\mu')^\alpha}\left[1 - \exp\left(-\sum_\nu \frac{C_O c^2 \mu\mu'}{C_{op} M \nu^{\beta-2}}\right)\right]. \tag{27}$$

Several cases are now possible. For $\beta > 3$ and $\alpha > 2$, it is easy to see that $\lim_{N\to\infty} \frac{E^{(P2)}}{E} = 1$ and higher edge multiplicities have essentially zero probability.

The situation is different if either condition is violated, since in this case $E^{(P2)} - E$ diverges and can become of the same order as E. For instance, we obtain for $\beta < 3, \alpha < 2$

$$E^{(P2)} - E \simeq \sum_{\mu,\mu'}^{\mu_{\max}} \frac{C_P^2 M^2}{(\mu\mu')^\alpha} \sum_{k\geq 2} \frac{(-1)^k}{k!}\left(\sum_\nu^{\nu_{\max}} \frac{C_O c^2 \mu\mu'}{C_{op} M \nu^{\beta-2}}\right)^k \tag{28}$$

$$\simeq \sum_{\mu,\mu'}^{\mu_{\max}} \frac{\text{const} \times M^2}{(\mu\mu')^\alpha} \sum_{k\geq 2} \frac{(-1)^k}{k!}(\text{const} \times \mu\mu' M^{3/\beta-2})^k \tag{29}$$

$$\simeq \sum_{k\geq 2} \text{const} \times \frac{(-1)^k}{k!} M^{2/\alpha + k(3/\beta + 2/\alpha - 2)}. \tag{30}$$

From the last formula, we see that the expected edge multiplicity $\frac{E^{(P2)}}{E} - 1$ can become positive for proper choices of α and β. We show that $\frac{E}{E^{(p2)}} < 1$ under the above assumptions. Since

$$E^{(P2)} = \sum_{\mu,\mu'} \sum_\nu \frac{C_O C_P^2 M}{C_{op}(\mu\mu')^{\alpha-1}\nu^{\beta-2}} \tag{31}$$

$$\simeq \text{const} \times M^{(1/\alpha)2(2-\alpha)+1+(1/\beta)(3-\beta)} \tag{32}$$

$$= \text{const} \times M^{4/\alpha + 3/\beta - 2} \tag{33}$$

and

$$E \simeq \sum_{k\geq 1} \text{const} \times \frac{(-1)^{k+1}}{k!} M^{2/\alpha + k(3/\beta + 2/\alpha - 2)}, \tag{34}$$

one gets

$$\frac{E}{E^{(P2)}} \simeq 1 - \sum_{k\geq 2} \text{const} \times \frac{(-1)^k}{k!} M^{2(k-1)/\alpha + 3(k-1)/\beta - 2k} \tag{35}$$

$$\simeq 1 - \text{const} \times M^{-2/\alpha - 3/\beta}[M^{2/\alpha + 3/\beta} - 1 + o(1)] \tag{36}$$

$$= 1 - \text{const} + o(1). \tag{37}$$

Since the involved constant is positive we get the desired result. A more careful analysis, which will be part of a forthcoming paper, shows that one also obtains a power law for the edge multiplicity, as observed in the simulations.

C. Random intersection graphs and the cameo principle

In this section, we give a possible explanation for the appearance of power laws in the size distribution. In most models of complex networks with power-law-like degree distributions, one assumes a kind of preferential attachment rule, as in the Albert and Barabási model [3]. This makes little sense in our framework. Instead we make use of the cameo principle, first formulated in [8].

Before giving an interpretation and motivation we briefly describe the formal setting. Assign to each project a positive,

φ-distributed random variable ω and to each organization a positive, ψ-distributed random variable σ (note that, in contrast to Sec. IV B, φ and ψ are not the size distributions). We assume φ and ψ to be supported on $(1,\infty)$ and monotone decreasing as ω and σ tend to infinity. We also make use of the notational simplifications $\varphi(P)=\varphi(\omega(P))$ and $\psi(O)=\psi(\sigma(O))$. On the bipartite graph, an edge between an organization O and a project P is then formed with probability

$$p_{o,p} := \frac{c_0}{\psi^{\alpha}(P)} \frac{1}{\sum_P \psi^{-\alpha}(P)} + \frac{c_1}{\varphi^{\beta}(O)} \frac{1}{\sum_O \varphi^{-\beta}(O)}, \tag{38}$$

where c_0 and c_1 are positive constants, the exponents α and β are in the interval $(0,1)$, and all edges are drawn independently of one another. We are interested in the properties of the corresponding random P and O graphs for typical realizations of the ω and σ variables. The word typical is here understood in the sense of the ergodic theorem, namely, we assume $\frac{1}{N}\Sigma_O \varphi^{-\beta}(O) \sim \int \varphi^{1-\beta} d\varphi =: C_0^{-1}$ and $\frac{1}{M}\Sigma_P \psi^{-\alpha}(P) \sim \int \psi^{1-\alpha} d\psi =: C_1^{-1}$, where N and M are the cardinalities of the O and P partitions and α and β are such that the integral is bounded. The above formula reduces then to

$$p_{o,p} := \frac{c_0 C_0}{M\psi^{\alpha}(P)} + \frac{c_1 C_1}{N\varphi^{\beta}(O)}. \tag{39}$$

The expected conditional size of a vertex is then given by

$$\mathbb{E}[|P||\psi(P)] = \frac{N c_0 C_0}{C_1 M \psi^{\alpha}(P)} + c_1 \tag{40}$$

and

$$\mathbb{E}[|O||\varphi(O)] = \frac{M c_1 C_1}{C_0 N \varphi^{\beta}(O)} + c_0. \tag{41}$$

The interpretation behind the special form of edge probability in Eq. (39) is the following. The ω and σ values describe a kind of attractivity property inherent to projects and organizations. Thinking in terms of a virtual project formation process either the final set of organizations belonging to a project P can join the project actively—in which case the σ value of P is important—or the organization more passively enters the project on the request of organizations already involved—in which case the attractivity ω of the the corresponding organization is important. The attractivity of an organization could, for instance, be related to its reputation, financial strength, or quality of earlier projects in which the organization was involved. The pairing probability is not directly based on the ω or σ values, but rather the relative frequency of the ω or σ values: the rarer a property, the more attractive it becomes. This is the essence of the cameo principle.

The parameters α and β can be seen as defining the propensity to follow the above rule; for $\alpha,\beta \to 0$ the rule is switched off and we recover a classical Erdös-Renyi intersection graph. In general the values of α and β are themselves quenched random variables with their own—usually unknown—distribution. As shown in [12], only the maximal α and β values matter for the resulting degree distribution of the graphs. We therefore restrict ourselves in the following to constant values.

Since the conditional expectations of the size values (Eqs. (40) and (41)) are proportional to $\varphi^{-\beta}$ and $\psi^{-\alpha}$, we have to estimate their induced distribution. It can be shown [13] that $z := \varphi^{-\beta}(\omega)$ is asymptotically distributed with density $z^{-[1+1/\alpha+o(1)]}$ when $\varphi(\omega)$ decays monotonically and faster than any power law to zero as $\omega \to \infty$. When $\varphi(\omega)$ is itself a power-law distribution with exponent γ, the resulting distribution for z will be $z^{-[1+1/\alpha-1/\alpha\gamma+o(1)]}$. Therefore, the induced distribution is always a power law and independent of the details of φ. Applying this result to our model, we obtain immediately a power-law distribution for the size distribution on the P and O graphs with exponents depending essentially only on α and β. Due to the edge independence in the model definition, the resulting degree distributions are again of power-law type. The cameo ansatz hence generates in a natural way a bipartite graph, where both projections admit two of the main features of the FP networks. Furthermore, we obtain a linear dependence of the mean triangle number $\triangle_k$ on the degree, as in Sec. IV A.

None of the models discussed in Sec. IV can reproduce scale-free distribution of the edge multiplicity with the same low exponent as observed in each of the FP networks. It will be interesting to see whether the inclusion of memory effects like the "my friends are your friends" principle [14] will change the picture.

V. CONCLUSIONS

In this paper, we have described research collaboration networks determined from research projects funded by the European Union. The networks are substantial in terms of size, complexity, and economic impact. We observed numerous characteristics known from other complex networks, including scale-free degree distribution, small diameter, and high clustering. Using a random intersection-graph model, we were able to reproduce many properties of the actual networks. The empirical and theoretical investigations together shed light on the properties of these complex networks, in particular that the EU-funded research and development networks match well with typical realizations of random graph models characterized by just four parameters: the size, edge number, exponent of project-projection degree distribution, and exponent of organization-projection degree distribution.

In terms of real-world interpretation, the present analysis yields three major insights. First, based on the fact that the size distribution of projects did not change significantly between the Framework Programs, any possible changes in project formation rules—which we do not know at this stage—did not affect the aggregate structure of the resulting research networks. Second, the fact that integration between collaborating organizations has increased over time, as measured by the average clustering coefficient, indicates that Europe has already been moving toward a more closely integrated European Research Area in the earlier Framework Programs. Finally, the fact that a sizable number of organizations collaborate more than once in each Framework Program shows that there appears to be a kind of robust backbone structure in place, which may constitute the core of the European Research Area.

In terms of application, the present results suggest a number of extensions. First, it is essential to learn more about the properties of the vertices in our networks. To what extent can they be characterized and classified? What kind of structural patterns emerge if we add this information? Second, we need to know more about the microstructure of the networks. In which areas are the networks highly clustered and where does this clustering come from? What kind of subgroups can be identified? Third, we need to learn more about where the observed distribution of edge multiplicity comes from. Finally, it would be desirable to explicitly include edge weights into the analysis, as actors who collaborate frequently are presumably more proximate to each other than actors who collaborate only once. This may significantly impact the structural features we are able to observe, as well as the conclusions we might draw concerning the link between network structure and function.

ACKNOWLEDGMENTS

We would like to acknowledge support from the Portuguese Fundação para a Ciência e a Tecnologia (Bolsa de Investigação SFRH/BPD/9417/2002 and FEDER/POCTI-SFA-1-219), ARC systems research (W4570000294-3), and the VW Stiftung (I/80496). We thank Ph. Blanchard and L. Streit for useful discussions and commentary. Portions of this work were done at the Vienna Thematic Institute for Complexity and Innovation, EXYSTENCE Network of Excellence: IST-2001-32802.

[1] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).

[2] B. Bollobás and J. Riordan, in *Handbook of Graphs and Networks* (Wiley-VCH, Berlin, 2003).

[3] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[4] M. Karonski, E. R. Scheinerman, and K. B. Singer-Cohen, Combinatorics, Probab. Comput. **8**, 131 (1999).

[5] K. Barker and H. Cameron, in *European Collaboration in Research and Development*, edited by Y. Caloghirou, N. S. Vonortas, and S. Ioannides (Edward Elgar, Cheltenham, U.K., 2004).

[6] CORDIS Projects Database—Advanced and Professional Database Search, http://dbs.cordis.lu/cordis-cgi/EI&CALLER =EIPROF_EN_PROJ&MODE=N&LANGUAGE =EN&DATABASE=PROJ.

[7] M. E. J. Newman, Phys. Rev. E **64**, 016131 (2001).

[8] Ph. Blanchard and T. Krueger, J. Stat. Phys. **114**, 5 (2004).

[9] M. Molloy and B. Reed, Random Struct. Algorithms **6**, 161180 (1995).

[10] B. Bollobás, Eur. J. Comb. **1**, 311316 (1980).

[11] Ph. Blanchard, A. Krüger, T. Krueger, and P. Martin, e-print physics/0505031.

[12] Ph. Blanchard, S. Fortunato, and T. Krueger, Phys. Rev. E **71**, 056114 (2005).

[13] Ph. Blanchard and T. Krueger, in *Extreme Events in Nature and Society*, The Frontier Collections (Springer, Berlin, 2006).

[14] Ph. Blanchard, T. Krueger, and A. Ruschhaupt, Phys. Rev. E **71**, 046139 (2005).