Maluck, J., Donner, R. V., Takayasu, H., Takayasu, M. (2017): Motif formation and industry specific topologies in the Japanese business firm network. - Journal of Statistical Mechanics: Theory & Experiment, 2017, 053404

DOI: 10.1088/1742-5468/aa6ddb

Motif formation and industry specific topologies in the Japanese Business Firm Network

Julian Maluck[1,2,3*], **Reik V. Donner**[3], **Hideki Takayasu**[1,4], **Misako Takayasu**[1]

[1] Institute of Innovative Research, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Yokohama 226-8502, Japan

[2] Humboldt University of Berlin, Department of Physics, 12489, Berlin, Germany

[3] Potsdam Institute for Climate Impact Research, 14473, Potsdam, Germany

[4] Sony Computer Science Laboratories, 3-14-13, Higashi-Gotanda, Shinagawa-ku, Tokyo, 141-0022, Japan

E-mail: *`Julian.Maluck@pik-potsdam.de`

13 April 2017

Abstract.

Motifs and roles are basic quantities for the characterization of interactions among 3-node subsets in complex networks. In this work, we investigate how the distribution of 3-node motifs can be influenced by modifying the rules of an evolving network model while keeping the statistics of simpler network characteristics, such as the link density and the degree distribution, invariant. We exemplify this problem for the special case of the Japanese Business Firm Network, where a well-studied and relatively simple yet realistic evolving network model is available, and compare the resulting motif distribution in the real-world and simulated networks. To better approximate the motif distribution of the real-world network in the model, we introduce both subgraph dependent and global additional rules. We find that a specific rule that allows only for the merging process between nodes with similar link directionality patterns reduces the observed excess of densely connected motifs with bidirectional links. Our study improves the mechanistic understanding of motif formation in evolving network models to better describe the characteristic features of real-world networks with a scale-free topology.

PACS numbers: 89.75.Hc, 89.65.Gh

Keywords: Network reconstruction, Inference in socio-economic system, Socio-economic networks

1. Introduction

In directed complex networks [1, 2] motifs are subgraphs that allow for a concise attribution of a node's functional role [3, 4]. These subgraphs are understood as essential building blocks of the networks and allow a detailed characterization and classification of the system under study in many disciplines, ranging from biology [5, 6] to socio-economic systems [7]. Previous work has provided a conceptual understanding of the mechanisms behind the formation of motifs in networks [8]. It has been shown that the connection patterns of nodes with high degrees have a particularly strong influence on the motif distribution [9]. From a dynamical perspective, motifs emerge in order to optimally exploit local stability characteristics of the network [10]. However, still little is known about the dynamical origins of the motif distributions in evolving complex networks with scale-free topology [11]. The aim of this study is to better understand the formation of motifs and to identify and test candidate mechanisms that could generically affect the formation process of motifs in an evolving network model.

Business and trade networks provide well-studied economic examples of directed networks that allow for detailed investigations of the aforementioned questions. In business networks, individual firms are interpreted as nodes [12]. If a business or trade relationship between two firms is reported, a link between the involved nodes is drawn. On a far more aggregated level, this system can be expressed as a trade network, where countries or industry sectors build the nodes of the network [13, 14, 15]. In both business and trade networks the nodes take specific roles within the supply chain as service or commodity providers and/or consumers. Both types of economic networks can be understood as arising from an aggregation of the supply chains for all individual products. On the level of international trade networks, industries have been identified to form characteristic trading patterns among different subnetworks [16]. On the firm level, previous empirical work has observed a strongly non-trivial motif distribution among Japanese business relations [17].

The business network of the Japanese economy shares some common topological features of many complex networks [18, 19]. These include the scale-free degree distribution that has been empirically observed in complex networks among a variety of different disciplines, ranging from social networks [20, 21, 22, 23], economics [24], biology [25] to information technology [26, 27]. The emergence of the scale-free property is often explained by preferential attachment [28, 29]. The comparatively small power-law exponent in the Japanese Business Firm Network is well modeled by a merging mechanism between firms that exhibits features that are statistically similar to processes in other fields in physics, such as the coagulation process in aerosols [30].

Due to the existence of universal features of complex networks and the straight-forward interpretation of a node's functional role within motifs, the Japanese Business Firm Network provides an interesting case for obtaining new conceptual knowledge about the formation of motifs in complex networks. The aim of this paper is to study possible candidate mechanisms that allow reproducing the empirical distribution of 3-node motifs in the Japanese Business Firm Network within an evolving network model. These mechanisms should leave the statistics of simpler network characteristics, such as the total number of nodes, link density and degree distribution as invariant as possible. Instead of providing a comprehensive review of the impacts of the investigated mechanisms on various topological network features at different levels of the network's structural organization, in this work we explicitly focus on obtaining an understanding on the formation of motifs.

We start our investigations by comparing some relevant topological characteristics of a previously suggested evolving network model [30] with those of the real-world business network in Japan. The utilized model is based on three main processes, node annihilation, node creation and merging of nodes, which are interpreted as bankruptcies, creation of newcomer firms, and a merging process between two firms, respectively. Firstly, we identify the main drawbacks in the description of motifs in the model. In a second step, we investigate the appearances of roles and motifs in different industries of the Japanese Business Firm Network and introduce an industry structure to the model. To influence the motif distribution in the model, we then evaluate the impact of different rules concerning the linking preferences of newcomer nodes. Furthermore, we monitor the appearance of motifs when modifying the rules for the merging process and when adding relinking possibilities to the model.

Our paper is organized as follows: In section 2, we describe the dataset (2.1) utilized in our study together with concepts to characterize the motif distributions (2.2) and the industry subgraphs (2.3). Subsequently, section 3 presents the existing basic evolving network model together with three modifications to improve the representation of the empirical motif distributions of the Japanese Business Firm Network. We then evaluate in section 4 the motif distributions in the real-world data and the original network model (4.1) and focus on distinctive topological characteristics within industry subgraphs (4.2) that are used to introduce subgraph specific rules in the network model (4.3). In section 4.4 we assess the impact of business preferences that can be expressed as global rules on the motif appearances. Section 5 summarizes our main findings.

2. Data and Methods

2.1. Dataset description

The investigated data encompasses business firm relationships that have been recorded by Teikoku Databank, one of Japan's largest corporate research providers. The database used for this study comprises company pairs that have reported an inter-company business relationship in the year 2009. Next to a company's industry attribution and its date of creation, the data contains information about the reporting date of an inter-company relationship.

Let $\mathcal{G} = (V, E)$ denote a directed unweighted network that consists of the set of nodes V and the set of links E, which is described by an adjacency matrix (a_{ij})

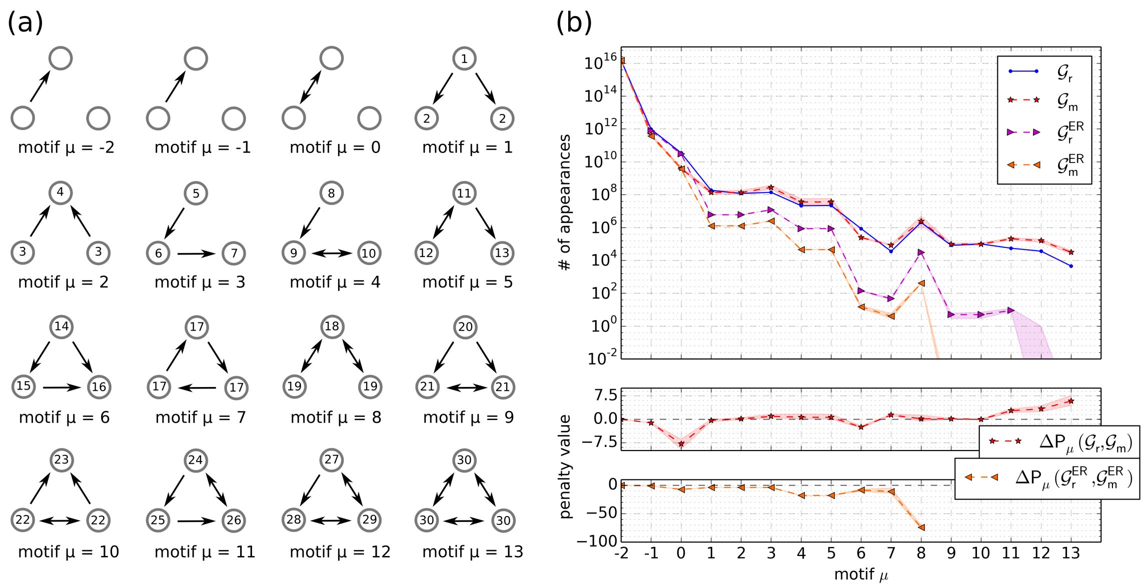

Figure 1. Distribution of motif appearances in different networks. (a) Overview of the motif patterns μ and roles R (encircled numbers) in directed networks. (b) Number of appearances of motif patterns in the real-world network $\mathcal{G}_r$ (blue solid), the original evolving network model $\mathcal{G}_m$ (red dashed) and in random graphs $\mathcal{G}_r^{ER}$ and $\mathcal{G}_\mu^{ER}$ with corresponding densities of uni- and bidirectional links as observed in the data (magenta dashed) and in the model (orange dashed), respectively. For the models, the median values are represented as markers and the upper (lower) quartiles are illustrated by the upper (lower) shaded boundary. The median and quartiles are estimated from the respective ensemble of model realizations. The two plots at the bottom show the penalty values for $\Delta P_\mu(\mathcal{G}_r, \mathcal{G}_m)$ (red dashed) and for the random graphs $\Delta P_\mu(\mathcal{G}_r^{ER}, \mathcal{G}_m^{ER})$ (orange dashed).

with entries $a_{ij} = 1$ if $(i,j) \in E$ and $a_{ij} = 0$ otherwise. We define the real-world Japanese Business Firm Network $\mathcal{G}_r = (V_r, E_r)$ by drawing a link $(i,j) \in E_r$ between two companies $v_i, v_j \in V_r$, if commodities or services have been exchanged in return for money. The link direction is defined to coincide with the direction of the monetary flow. As the available data does not contain quantitative information on the value of the flow, in this work we will exclusively consider such binary networks. The resulting real-world network consists of $|V_r| = 446,108$ companies with $|E_r| = 2,471,689$ links. The results of this study have been cross-checked for consistency with the available data for the firm networks from 2010 to 2015. As the main results do not change significantly with the years, we restrict our following discussions to the results for the network of 2009.

2.2. Motif and role comparisons

Motifs are defined as small induced subgraphs of the network. In directed networks, 13 connected motif patterns (denoted with index μ) with 3 nodes can be distinguished. Within these motifs, the individual nodes occupy one of 30 characteristic functional roles (index R). An overview of the motif patterns and roles is presented in figure 1(a).

If one node is a member of various motifs in the graph, this node can simultaneously take different roles.

In previous studies [17] a non-trivial motif distribution of connected subgraphs has been observed in the Japanese Business Firm Network. The results have been obtained by comparing the appearances of motifs in the real-world network with expected appearances in randomized graphs. These surrogates were generated by switching links such that each node keeps its number of incoming and outgoing links [8]. This is a useful approach to characterize the motif appearance of real-world networks with respect to a randomized graph with identical degree sequence as the original model.

In contrast, in this study we are interested in the total differences between the motif distributions in two networks $\mathcal{G}_x$ and $\mathcal{G}_y$, which partially, but not exclusively arise from differences in the local in- and out-degrees. Consequently, we directly compare the absolute number of motif appearances rather than their normalized deviations from the case of a randomized null model. This avoids the problem that differences in the topological characteristics of the surrogates of the two networks will not be tracked, if the networks $\mathcal{G}_x$ and $\mathcal{G}_y$ are compared to their respective surrogates. To allow for a comparison of the absolute number of motif appearances, we ensure that the number of nodes in the ensemble average of the model networks complies with the number of nodes in $\mathcal{G}_r$.

With x_μ denoting the number of appearances of motif pattern μ in a graph $\mathcal{G}_x$ and y_μ the respective number in another graph $\mathcal{G}_y$, we define the *penalty value* $\Delta P_\mu(\mathcal{G}_x, \mathcal{G}_y)$ as

$$\Delta P_\mu(\mathcal{G}_x, \mathcal{G}_y) := \frac{x_\mu - y_\mu}{\min(x_\mu, y_\mu)} \ . \tag{1}$$

The difference between the appearances of motif μ is normalized with respect to the graph with the smaller number of appearances. This allows for a meaningful comparison between different motif patterns with different magnitudes in appearance. Equation (1) is anti-symmetric with respect to the exchange of graphs, following $\Delta P_\mu(\mathcal{G}_x, \mathcal{G}_y) = -\Delta P_\mu(\mathcal{G}_y, \mathcal{G}_x)$. To obtain the penalty value for an ensemble of model realizations, we first average the motif frequencies of all ensemble members and then apply equation (1).

2.3. Industry subgraphs

Motif distributions and linking patterns are known to show industry specific characteristics in macroeconomic networks [16, 31]. To investigate how these characteristics are present in the microeconomic network of business firms, we assess how the industry structure in the Japanese Business Firm Network provides information about the composition of the network's topology and the motif distribution. Let $G_\zeta = (V_\zeta, E_\zeta)$ denote the subgraph that is induced by the subset of nodes $V_\zeta \subset V_r$. Here, V_ζ contains all nodes of $\mathcal{G}_r$ that belong to an industry ζ.

To analyze differences in the topology of the individual subgraphs, we start with examining the degree distribution, the link density and the reciprocity r_ζ of the individual industries' induced subgraphs. Since the analyzed networks do not contain self-loops and as we are interested in both the link density and the reciprocity separately [32], we utilize the standard definition of the reciprocity,

$$r_\zeta = \frac{\sum_{i\neq j} a_{ij}^{(\zeta)} a_{ji}^{(\zeta)}}{\sum_{i\neq j} a_{ij}^{(\zeta)}} , \tag{2}$$

where the $a_{ij}^{(\zeta)}$ denote entries of the adjacency matrix of subgraph G_ζ. This definition describes the probability of the existence of an oppositely directed link when one link in the network is chosen at random.

Furthermore, we investigate the characteristic linking patterns of industries by counting the appearance of roles in motifs. To assess an industry's role characteristic, we aggregate the appearance of role R in industry ζ by considering all nodes of this industry, obtaining the number of appearances in the industry

$$N_\zeta^R = \sum_{i \in G_\zeta} \nu_i^R . \tag{3}$$

The number of appearances of role R for one node i is denoted by ν_i^R. Here, motifs of the full network $\mathcal{G}$ that may extend over several industries are taken into account. In order to compare the role characteristic for industries of different size, we normalize N_ζ^R with respect to the full number of role appearances in this industry,

$$n_\zeta^R := \frac{N_\zeta^R}{\sum_{R_i} N_\zeta^{R_i}} . \tag{4}$$

The associated z-score, defined as

$$z_\zeta^R := \frac{n_\zeta^R - E_\zeta(n^R)}{\sigma_\zeta(n^R)} , \tag{5}$$

quantifies to which extent a role R is characteristic for industry ζ. Here, $E_\zeta(n^R)$ $[\sigma_\zeta(n^R)]$ denotes the mean value [standard deviation] of the normalized appearances among the industries. A high value of z_ζ^R implies a relatively frequent appearance of role R in industry ζ.

3. Evolving network model

3.1. Basic model

In previous work, it has been shown that some key topological features of the real-world Japanese Business Firm Network $\mathcal{G}_r$ can be well approximated by an evolving network model $\mathcal{G}_m$ that incorporates a merging process of firms [30]. The development of this model has been originally motivated by the observation of similar statistical characteristics in the mass distribution of aerosols [33]. In each evolution step of this model, a randomly selected node is either annihilated (with probability a), newly created (probability b) or merged with another existing node in the network (probability $c = 1 - a - b$). When a new node is created one incoming and one outgoing link are attached to the new node. In order to reproduce the observed scale-free degree distribution of the real-world network, the selection criteria for both link creation and the merging process follow a preferential attachment rule [34]. Previous studies have demonstrated that the merging activity of firms in the real economy follows waves of market sentiment and depends on many factors, for example the size and the stock market performance of the involved firms [35, 36]. As a simplification, we do not consider these factors here and utilize the simple assumption of preferential attachment to reproduce the basic statistical characteristics of the real-world network.

The probability $\Pi_{j\to i}$ that node j is connected to (or merged with) node i reads

$$\Pi_{j\to i} = \frac{k_i + 1}{\sum_l (k_l + 1)} \, . \tag{6}$$

Thus, a node i with a high total degree k_i (i.e. the number of incoming plus outgoing links) is more likely to be chosen as business partner. With the choice of equation (6) we ensure that isolated nodes with $k_i = 0$ can still be chosen as new business partners.

As we will show in section 4, there are marked differences between the motif distributions obtained from the basic evolving network model as described above and those of the empirical Japanese Business Firm Network. In order to resolve these discrepancies, in the following we utilize the characteristics of the industry subgraphs in the real-world network $\mathcal{G}_r$ to deduce meaningful modifications of the model. Thus, we introduce an industry structure to the model and assess the impact of characteristic linking and merging preferences between firms on the appearance of motifs in the network. Specifically, we will consider three types of modifications that are described in the following.

3.2. Introduction of industry structure

Motivated by empirical findings that are obtained by utilizing the methods introduced in section 2.3 and detailed later in section 4.2, we consider an additional industry structure as part of the network model. At the initiation of the network evolution, we randomly assign an industry to each node. The probability of a node to be associated with industry ζ is estimated from $\mathcal{G}_r$ as $|V_\zeta|/|V_{r'}|$, where $|V_{r'}|$ denotes the total number of firms that belong to either of the following five main industries of the network (cf. section 4.2): construction, manufacturing, whole sales, transport & communication and service. We assign to newcomer nodes the industry with the largest population deficit that has arisen from previous annihilation and merging processes in comparison with the real-world benchmark network.

In the original model, a new node is created with exactly one incoming and outgoing link each. Here, we modify the connection pattern of newcomers and introduce two characteristic connection patterns for each industry of which one is randomly chosen at node creation with equal probability. This procedure represents a simplification of the patterns observed in the real-world network as will be discussed in section 4.3. However, by restricting the modification to two linking patterns with equal probability, a clear understanding of the influence of linking preferences of newcomer firms on the resulting motif distribution can be obtained. Furthermore, we reduce the risk of over-fitting the model by minimizing the introduction of additional parameters. The choice of the industry specific connection pattern is motivated by the results of the subgraph analysis of the real-world network $\mathcal{G}_r$, for which the obtained specific connection patterns for each industry will be presented in section 4.3. We refer to the model that includes the industry structure as $\mathcal{G}_{ind}$ in the remainder of this study.

3.3. Δk-rule for the merging process

In section 4.1, we will report and discuss an observed excess frequency of motifs with bidirectional links in the model $\mathcal{G}_m$ as compared to $\mathcal{G}_r$. To account for this effect, we designed a rule to reduce this overshoot in the model. Bidirectional links occur in the merging process when two companies merge that share the same business partner. When two firms serve as a supplier and consumer to the shared partner before merging, respectively, a bidirectional link will be established in the merging process. Thus, an intuitive rule to reduce the appearance of motifs with bidirectional links is based on the local degree properties of each node. We denote the difference of a node's in-degree and out-degree as $\Delta k_i = k_i^{in} - k_i^{out}$. Following the above argument, we expect a reduction of bidirectional links if merging processes between firms i and j with $\Delta k_i \cdot \Delta k_j < 0$ are forbidden. This rule, which we refer to as the Δk-rule, can be interpreted such that two firms with similar shares of their input/output allocations are more likely to merge. With the production function describing a company's output depending on its input, two firms with similar production functions will exhibit the same sign in Δk_i, allowing for a merging process. In this paper, we indicate network models $\mathcal{G}_\bullet$ that include the Δk-rule with an additional subscript Δk.

3.4. Substitution of trading partners

Finally, we investigate a third modification of the model that allows companies to substitute their trading partners. Hence, we introduce a relinking process with probability d in the evolution step. The probabilities of the processes during network evolution are then adjusted to satisfy $a' + b' + c' + d = 1$. Specifically, the probabilities of the annihilation, creation and merging processes are modified such that the relative ratios between the probabilities of the processes stay constant, i.e. $a' = (1-d) \cdot a$ and analogous for b' and c'. During relinking, a company (node) i and one link are chosen at random. We then consider this company substituting its business relation in favor of a company that shares a common business partner. Thus, the new business partner l of node i is found by following a path from node i to another node of length 2. With the symmetrized adjacency matrix with entries $a'_{ij} = a'_{ji} = \max(a_{ij}, a_{ji})$, the probability that the link is reconnected to node l is then given by

$$\Pi_{i,l} = \frac{(\sum_j a'_{ij} a'_{jl})(1 - a'_{il})}{\sum_m (\sum_j a'_{ij} a'_{jm})(1 - a'_{im})} \,. \tag{7}$$

Equation (7) accounts for the assumption that more common business relations make a reconnection to node l more likely. Models $\mathcal{G}_\bullet$ that include the relinking process are denoted with the additional subscript l.

4. Results and discussion

4.1. Motif appearances in the real-world network and basic model

Previous work has shown that the model $\mathcal{G}_m$ with a choice of probabilities $a = 0.2$, $b = 0.5$ and $c = 0.3$ (for the annihilation, creation and merging process, respectively) approximately reproduces the scale-free degree distribution from the real-world network, following a power law $\propto k^{-\alpha}$ with an exponent of $\alpha \approx 1.4$ [12]. Despite changes in the number of firms and links, it has been concluded that some basic characteristics such as the degree distribution are robust in the business firm network of Japan from 1994 to 2014 [12]. In this work, we simulate an ensemble of 100 networks with approximately the same number of nodes $|V_m|$ in the model as in $\mathcal{G}_r$. From the different realizations, we obtain a network ensemble with $|V_m| = 450,000 \pm 6,000$ nodes and $|E_m| = 1,090,000 \pm 20,000$ links (denoting the ensemble average and standard deviation, respectively). Note that due to the probabilistic nature of the network model, it is not possible to fix both the number of nodes and the link density at exactly the desired values corresponding to those of the real-world network.

Figure 1(b) illustrates the motif distribution for the real-world network $\mathcal{G}_r$ and the basic evolving network model $\mathcal{G}_m$. Note that many patterns of connected motifs are more frequently observed in $\mathcal{G}_m$ despite its lower link density ($\rho_m < \rho_r$). To further quantify the impact of the link density difference, we construct an Erdős-Rényi type random graph ensemble $\mathcal{G}_r^{ER}$ such that the link densities for both uni- and bidirectional links correspond to the respective densities in the real-world network. This is achieved by drawing uni- and bidirectional links with the respective probability estimated from the real-world network separately. $\mathcal{G}_m^{ER}$ is similarly defined for the modeled network. We simulate 100 realizations for the Erdős-Rényi type random graphs. As shown in figure 1(b) the penalty values $\Delta P_\mu(\mathcal{G}_r, \mathcal{G}_m)$ are in general smaller than the corresponding values $\Delta P_\mu(\mathcal{G}_r^{ER}, \mathcal{G}_m^{ER})$ between the random networks with identical link density. This observation arises from the similarity in the degree distribution between data and model, both following a power-law $\propto k^{-\alpha}$ with almost identical exponent. In random graphs, the Poisson distributions show different values of their characteristic parameter which are uniquely determined by the respective link density. For sparsely connected motif patterns, the evolving network model offers a good description of the motif distribution of the data. However, the model considerably overestimates the appearances of dense motifs with bidirectional links such as $\mu = 11$ to 13, especially $\Delta P_{13}(\mathcal{G}_r, \mathcal{G}_m) = 5.8^{+1.8}_{-1.4}$. Here, the central penalty value refers to the median of the ensemble. The upper (lower) error bands denote the differences of the upper (lower) ensemble quartiles of ΔP_μ from the corresponding median. As we will demonstrate in section 4.4, the large value of

ΔP_{13} can be attributed to the creation of bidirectional links during the merging process in the model.

4.2. Industry subgraphs in the business network

In the real-world network, 95.6 % of all firms belong to either of the five industry subgraphs with highest cardinality: whole sales (G_W, $|V_W|/|V_r| = 0.329$), construction (G_C, 0.22), manufacturing (G_M, 0.215), services (G_S, 0.145) and transport & communication (G_T, 0.047). As only a small part of the network consists of nodes from neither of these subgraphs, we focus in this study on these main industries. Figure 2(a) shows the degree distributions of the subgraphs. From visual inspection, we observe that in most industries this distribution resembles a power-law with an exponent that is similar to the exponent of $\alpha \approx 1.4$ in the full network. In the construction sector, however, the distribution does not obey a power-law due to the lack of middle-sized firms. This underlines the uniqueness in the establishment of business relations between construction firms in Japan, which is supported by findings from previous studies [37].

We further analyze the characteristic linking patterns of industries by counting the appearances of roles in motifs. In business networks, a node's role is related to a company's function in the supply chain of a final product. For example, the pattern of motif $\mu = 3$ (see figure 1(a)) arises from a supply chain where the firm with role $R = 6$ serves as an intermediate component manufacturer for some end product. Note that individual supply chains cannot be obtained from the network at the considered aggregation level of the data. Nevertheless, the role of a firm can provide meaningful interpretations. For example, if a firm specializes on the production of intermediate components, it is likely that role $R = 6$ is attributed to that firm in many motifs.

The obtained z-scores, determined as described in equation (5), are presented in figure 2(b). We observe that manufacturing companies are more likely to occupy roles as a provider of intermediate goods ($R = 6$) and roles with a bidirectional link to one partner. Business firms in the service sector, however, are more prominent in roles with incoming links ($R = 2, 7$ and 13). Figure 3 illustrates excerpts of subgraphs that consist of example business firms with a characteristic role for their industry. Here, the representative firm in the construction sector (figure 3(a)) occupies role $R = 1$ by consuming goods and services from other firms. The providers are primarily firms from the construction or whole sales sector. In the example from the manufacturing industry (figure 3(b)), the considered firm exhibits mainly outgoing links to the whole sales sector. This indicates the requirement of inputs for production delivered by that sector. The large number of incoming links from other manufacturing companies hint towards the firm's role as a manufacturer of intermediate goods. In the whole sales sector (figure 3(c)) the selected firm with role $R = 18$ has established bidirectional relationships with other whole sales and manufacturing businesses. A different pattern is observed for the transport & communication and service sectors. Here, nodes with characteristic roles often exhibit a low degree but are connected to hubs in other industries. In

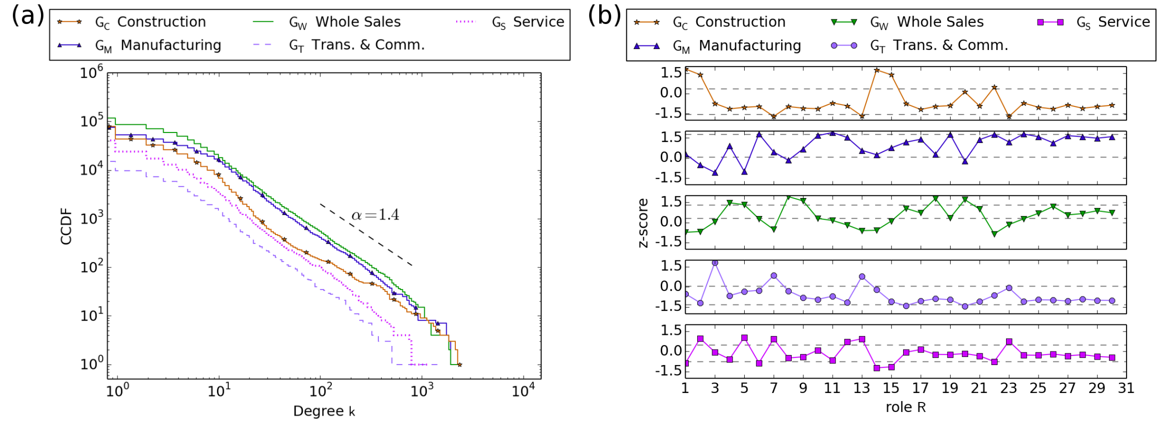

Figure 2. Topological properties of the industry subgraphs. (a) Complementary cumulative distribution function of the total degree k_i in the induced subgraphs G_ζ for different industry categories ζ. (b) Role analysis of industries: z-scores as defined in equation (5). The values of $E(z_\zeta) \pm \sigma(z_\zeta)$ for each industry ζ are indicated by gray lines.

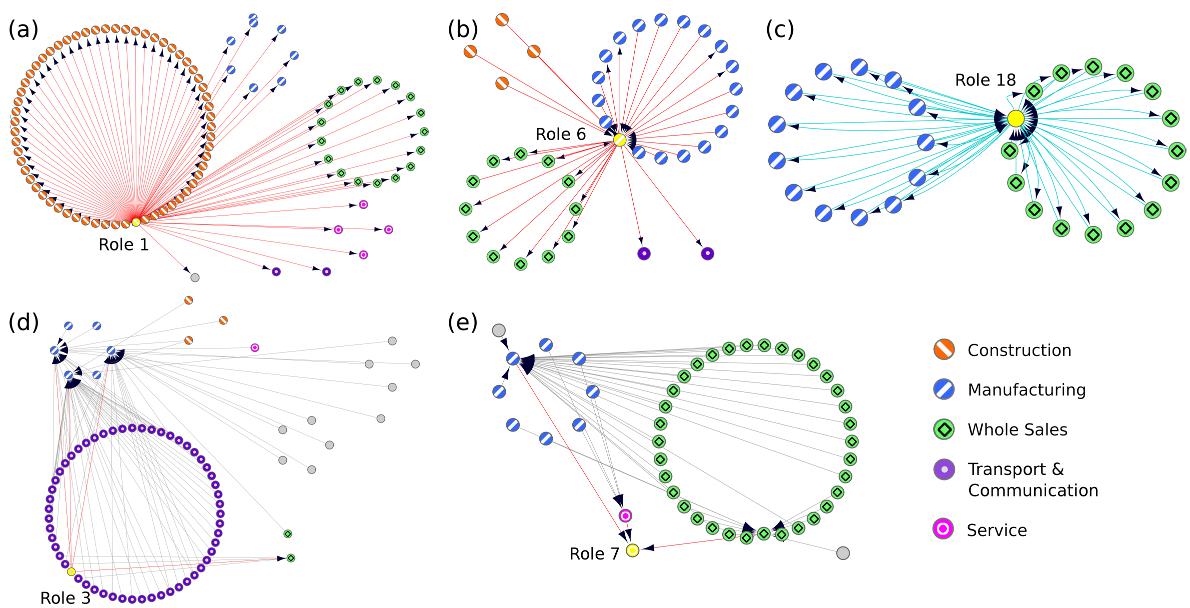

Figure 3. Network excerpts around exemplary nodes with characteristic industry roles. The exemplary nodes from the construction (a), manufacturing (b), whole sales (c), transport & communication (d) and service (e) sectors are marked yellow, respectively. Unidirectional (bidirectional) links connected to the selected nodes are plotted in red (light blue). Gray edges depict unidirectional links that are not connected to the selected nodes.

figure 3(d), many firms from the transport & communication sector consume goods or services from the same manufacturing partner. Finally, in figure 3(e), the selected firm provides services to a manufacturing business that has many clients in the whole sales sector. A summary of the characteristic topological features in the industry-induced subgraphs, including the link density ρ_ζ and the reciprocity r_ζ, is presented in table 1.

Table 1. Overview on some topological characteristics in the key industries of the Japanese economy and illustration of the introduced linking pattern in the modified model $\mathcal{G}_{ind}$. Depending on a newcomer's industry, one of the two patterns is chosen at random with equal probability at node creation.

Subgraph	r_ζ	ρ_ζ	characteristics of dominant roles	pattern in $\mathcal{G}_{ind}$ (I)	(II)
G_C Construction	0.04	$1.9 \cdot 10^{-5}$	both incoming and outgoing links	$\rightarrow\bullet\rightarrow$	$\rightarrow\bullet$
G_M Manufacturing	0.11	$3.3 \cdot 10^{-5}$	intermediate/component manufacturer	$\rightarrow\bullet\rightarrow$	$\leftrightarrow\bullet\leftarrow$
G_W Whole Sales	0.06	$2.0 \cdot 10^{-5}$	more outgoing than incoming links	$\leftarrow\bullet\rightarrow$	$\rightarrow\bullet$
G_T Transport & Comm.	0.10	$9.0 \cdot 10^{-5}$	unidirectional links	$\rightarrow\bullet\leftarrow$	$\leftarrow\bullet\rightarrow$
G_S Service	0.07	$2.2 \cdot 10^{-5}$	incoming links	$\rightarrow\bullet\leftarrow$	$\rightarrow\bullet$

4.3. Impact of the industry structure on the motif appearances

With the results presented above, we next introduce an industry structure to the model and modify the linking pattern at the creation process of a new node as described in section 3.2.

The considered linking patterns in $\mathcal{G}_{ind}$ and the results of the subgraph analysis of $\mathcal{G}_r$ are summarized in table 1. The comparatively high value of the reciprocity r in the manufacturing sector G_M indicates a higher density of bidirectional links as compared to most other industries. Therefore, we introduce the possibility of a bidirectional link for newcomer nodes from the manufacturing sector. We observe that $z^r_M > E(z_M) + \sigma(z_M)$ for $R = 6, 11, 19$ and 24 (cf. figure 2). Except for $R = 19$ these roles describe the production of intermediate goods. Thus, we keep the pattern of one incoming and one outgoing link for manufacturing nodes. In the transport & communication industry G_T the reciprocity also shows a comparatively high value. However, we do not include a bidirectional link for G_T due to the infrequent appearance of roles with these links (i.e. roles $R = 9$ - 12 and $R = 26$ - 30, see figure 2). This illustrates the importance of the connection to high-degree firms in other industries, as shown in figure 3(d). As the roles $R = 3, 7, 13$ with $z^r_T > E(z_T) + \sigma(z_T)$ exhibit either incoming or outgoing links, we introduce the patterns with two links of identical directions (either incoming or outgoing) to the node for G_T. Motivated by the lower link density ρ_ζ in G_C, G_W and G_S, we introduce the pattern with a single unidirectional link only for newcomer nodes of these industries. As most motifs μ appear more frequently in $\mathcal{G}_m$ than in $\mathcal{G}_r$ (cf. figure 1), we do not consider the possibility of adding more than two links to newcomers. In the whole sales sector G_W, we observe more outgoing than incoming links for the characteristic roles $R = 4, 5, 8, 9, 18$ and 20. On the other hand, we see a strong dominance of roles $R = 2, 5, 7, 12$ and 13 with incoming links in the service sector G_S. In G_C both incoming and outgoing links are almost equally present in the roles with $z_C > E(z_C) + \sigma(z_C)$. We consider these results in the directionality of links in the selected patterns as summarized in table 1.

To compare the introduced connection patterns with the empirical data, we analyze the directionality distribution of newly established firms in the Japanese economy from

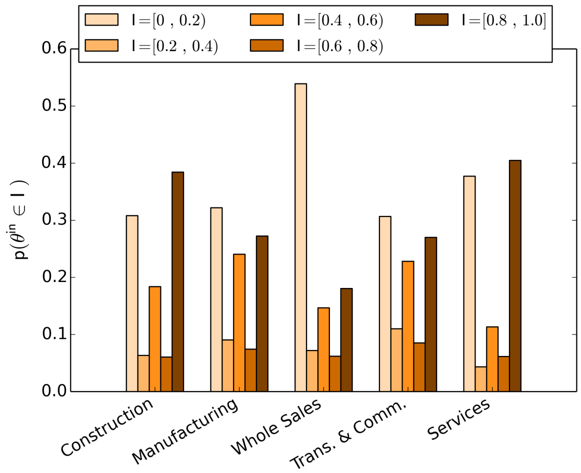

Figure 4. Distribution of the fraction of incoming initial links θ^{in} for key industries in the Japanese Business Firm Network. The bars indicate the probabilities $p(\theta^{in} \in I)$ of values in the intervals I defined as described in the legend.

$\mathcal{G}_r$. For this purpose, we define the fraction of incoming initial links θ_i^{in} for node i as

$$\theta_i^{in} = \frac{\kappa_i^{in}}{\kappa_i} \ . \tag{8}$$

Here, κ_i denotes the number of initially established links of company i and κ_i^{in} describes the number of incoming links among them. Motivated by a comparison of the real-world data with the original evolving network model we consider $\kappa_i = 2$. There are, however, firms with more than 2 links, that have been reported in the same month as the establishment of the first link. In these cases the first two links cannot be uniquely identified and we set the value of $\kappa_i > 2$ such that all reported links of this month are considered. In cases when the first 2 links of a newcomer firm can be obtained from the data, the statistic of θ^{in} in equation (8) allows for a direct evaluation of the linking pattern of newcomers in the real-world network that can be compared with the introduced linking pattern (I) of $\mathcal{G}_{ind}$ in table 1. If $\kappa_i > 2$, the method still provides insights to the question if the first links of newcomers are predominantly incoming or outgoing.

The distribution of θ^{in} is illustrated in figure 4. We observe a strong tendency of firms in the whole sales sector to firstly establish outgoing links to other companies. Firms in this industry buy the products first before further distributing them among other companies or final consumers. In the construction and service sectors, new participants in the market tend to start as providers, establishing incoming links first. Although the observed connection patterns are less distinctive than in the model, these findings are in line with the introduced simplified pattern scheme in table 1.

The impact of the introduction of the industry structure to the model as described above can be seen in figure 5. We observe that the industry structure improves the description of motifs $\mu = 0$, 3, 7, 11, 12, and 13, whereas the absolute penalty value

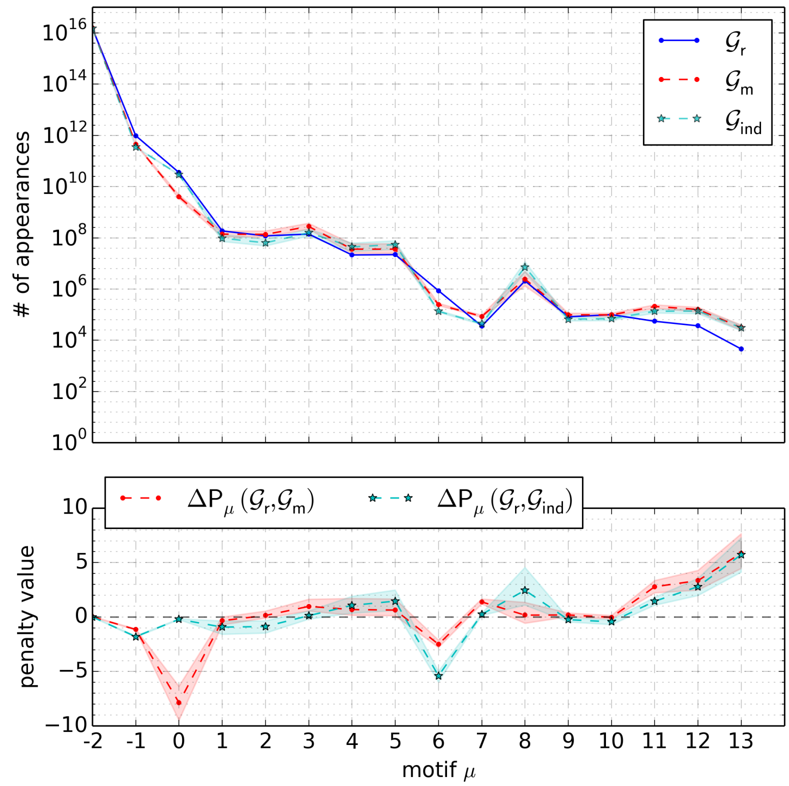

Figure 5. Impact of model $\mathcal{G}_{\mathbf{ind}}$ on the motif distributions in the full network. Top: Absolute frequencies of motif appearances in the real-world network $\mathcal{G}_r$ (solid blue), the model $\mathcal{G}_m$ (red) and the modified model with and industry structure $\mathcal{G}_{ind}$ (light blue). Bottom: Penalty values ΔP_μ as defined in equation (1) for two model variants with respect to the real network. For the models, the median (markers) and the respective upper and lower quartile (shaded boundaries) of the ensemble realizations are shown.

increases for motifs $\mu = 1$, 2, 4, 5, 6 and 8. However, the improvements outweigh the negative effects (see table 2). Nevertheless, despite a positive impact on the densely connected motifs $\mu = 11$, 12 and 13, further modifications of the model are required to more accurately describe the motif distribution of the real-world network.

4.4. Impacts of merging preferences and substitution of links on the motif appearances

From the findings discussed above, we deduce that a reduction of the observed excess frequency of densely connected motifs in the model as compared to $\mathcal{G}_r$ requires an additional modification of the merging rule. Specifically, this excess cannot be simply explained by the introduction of an industry structure to the model alone (see figure 5). Therefore, we utilize in the following the Δk-rule (cf. section 3.3). Thus, only pairs of firms with similarities in their input/output allocations merge in this model variant. We apply this rule in addition to the introduction of the industry structure and investigate the results of the resulting model $\mathcal{G}_{ind,\Delta k}$.

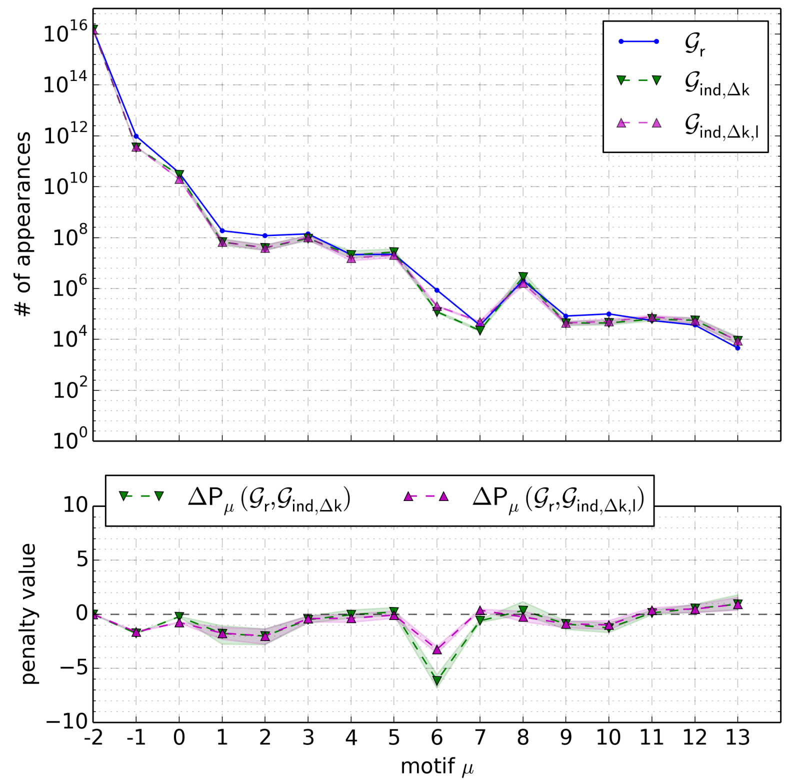

Figure 6. Impact of the Δk-rule on the motif distributions in the full network. As in figure 5 for two further model variants $\mathcal{G}_{ind,\Delta k}$ and $\mathcal{G}_{ind,\Delta k,l}$ involving the Δk-rule.

Figure 6 shows the total numbers of motif appearances (top) and the penalty values ΔP_μ (bottom) for the modified model $\mathcal{G}_{ind,\Delta k}$ that includes both the industry structure and the Δk-rule. We observe that the introduction of the Δk-rule has a substantial impact on motifs with bidirectional links, greatly reducing the number of appearances of motifs $\mu = 8, 11, 12$ and 13. This leads to a much better description of the data in these motif patterns, with a penalty value of $\Delta P_{13}(\mathcal{G}_r, \mathcal{G}_{ind,\Delta k}) = 0.9^{+0.8}_{-0.5}$ in motif $\mu = 13$. This improvement of the model leads to the hypothesis that pairs of firms that allocate their input/output similarly show a higher probability to merge. A more detailed investigation of this hypothesis is subject to future studies in order to complement the literature on drivers of the merging of firms [35, 36].

As shown in section 4.3 the introduction of the industry sector results in a better description, especially of unconnected motifs. This is mainly achieved by introducing the bidirectional link pattern to manufacturing firms. For the sparsely connected motifs $\mu = 1$ to 5 the models $\mathcal{G}_m$ and $\mathcal{G}_{ind}$ describe the motif appearance as observed in the real-world network almost equally well (cf. figure 5). In turn, this rule provides a more accurate description of unconnected motifs with bidirectional links ($\mu = 0$) and reduces the deficit of densely connected motifs $\mu = 11$ to 13. However, for the fully connected

Table 2. Sum of the penalty values for all motifs in the investigated model variants.

model	$\mathcal{G}_m$	$\mathcal{G}_{ind}$	$\mathcal{G}_{ind,\Delta k}$	$\mathcal{G}_{ind,\Delta k,l}$	$\mathcal{G}_{\Delta k}$	$\mathcal{G}_l$	$\mathcal{G}_{\Delta k,l}$	$\mathcal{G}_{ind,l}$
$\lvert\Delta P(\mathcal{G}_r, \mathcal{G}_\bullet)\rvert$	28^{+11}_{-6}	25^{+10}_{-8}	17^{+7}_{-4}	14^{+5}_{-4}	33^{+15}_{-10}	29^{+9}_{-5}	32^{+13}_{-9}	21^{+8}_{-6}

motif with unidirectional links ($\mu = 6$) the introduction of the industry structure results in a less accurate modeling of the real-world network. This can be attributed to the introduction of nodes in three industries that receive only one link at creation.

Furthermore, the deficit of appearances of motif $\mu = 6$ can be explained by the lack of relinking possibilities for companies in the models $\mathcal{G}_m, \mathcal{G}_{ind}$, and $\mathcal{G}_{ind,\Delta k}$. As described in section 3.4 and equation (7), we therefore finally investigate the impact of additional relinking possibilities on the motif distribution by considering substitution opportunities for companies according to their preference towards common business partners in the model variant $\mathcal{G}_{ind,\Delta k,l}$. Thus, the abundances of unidirectional fully connected motifs are increased.

To compare the description of the motif distributions among all variants of the evolving network model, we calculated the sum $|\Delta P| := \sum_\mu |\Delta P_\mu|$ over the 16 motif patterns μ for all model modifications in table 2. The table includes all combinations of the discussed modifications. We see that the Δk-rule and the introduction of relinking possibilities alone do not decrease the penalty value $|\Delta P|$ in comparison with the original model $\mathcal{G}_m$. This can be attributed to the fact that while these models improve the descriptions of motifs they are defined for (i.e. $\mu =$11, 12, 13 for Δk and $\mu = 6$ for l), the unconnected motif with a bidirectional link ($\mu = 0$) is still badly described without including the industry structure in the model. Thus, the additional industry structure improves the description of the motif distribution in the model markedly. From table 2 and figures 5 and 6, we see that, consequentially, the model $\mathcal{G}_{ind,\Delta k,l}$ offers the best description of the motif distribution.

To find a suitable value for the relinking probability d in model $\mathcal{G}_{ind,\Delta k,l}$, we investigated the impact on the motif distribution for different values of d. The possibility to substitute business relations improves the description of motif pattern $\mu = 6$ in the model. A high value of d results in a high frequency of this pattern (not shown). However, the modification l also increases the frequency of pattern $\mu = 7$ and decreases the frequency of $\mu = 0$ and 8 (not shown). With these opposing effects on the motif distribution, we observe the lowest sum of the penalty values for $d = 0.3$. At this probability the largest effect on motif $\mu = 6$ is obtained, while leaving the impacts on other motif patterns relatively small. Compared to $\mathcal{G}_{ind,\Delta k}$, the penalty value $|\Delta P_6|$ considerably decreases without increasing $|\Delta P_\mu|$ for $\mu \neq 6$. The median value of $\sum_\mu |\Delta P_\mu|$ decreases from 17^{+7}_{-4} for $\mathcal{G}_{ind,\Delta k}$ to 14^{+5}_{-4} for $\mathcal{G}_{ind,\Delta k,l}$.

From this observation, we conclude that substitutions of business relations play an important role in the formation of motifs. In particular, our results imply that the representation of the motif distribution in the model improves when business relations are relinked by firms in favor of partners that already exhibit many common business partners. We consider a study on empirical evidences for this hypothesis concerning the relinking process as important subject for future research.

5. Conclusion

In this study we have studied different strategies to influence the motif distribution in an evolving network model to better understand the formation process of motifs in the Japanese Business Firm Network. While the original model provides a good description of many topological properties of the real-world network, it overestimates the appearance of densely connected motifs with bidirectional links. To better understand the origin of motif appearances, we have analyzed the topology of individual industry subgraphs in the real-world network and introduced an industry structure to the model. Unlike other industries, the construction sector in Japan follows a unique degree distribution that does not obey a power-law due to the lack of middle-sized firms. The performed role analysis with respect to industries demonstrated that roles with bidirectional links appear predominantly in the manufacturing sector.

We have introduced subgraph- and industry-specific linking patterns at node creation in the evolving network model. This influences the appearance of sparsely connected motifs within the individual subgraphs. However, the impact on densely connected motifs with bidirectional links is small. To better describe the appearance of densely connected motifs in the real-world network, we modified the merging process of nodes in the model by introducing the Δk-rule. This rule states that nodes that exhibit higher in-degree than out-degree do not merge with nodes that have more outgoing than incoming links. In the context of business networks this rule is interpreted such that firms which predominantly provide goods or services to other companies do not merge with firms that predominantly exhibit the position of a customer. With the Δk-rule the frequencies of densely connected motifs with bidirectional links in the model are reduced, improving the reproduction of the real-world network properties in the model. Furthermore, we have shown that by introducing additional possibilities to reconnect links from a node preferably to another node with many common neighbors, the deficit of appearances of completely unidirectionally connected motifs in the model is reduced.

Our results provide both empirical and conceptual insights into the mechanisms of motif formation in an evolving network model. We have modified the existing basic model to improve the description of the motif distribution by achieving a better agreement with the real-world business network. In addition, our findings provide information about the selection preferences of business partners in the creation and merging process of firms. We have established the hypothesis that business firms tend to merge with companies that exhibit similarities in their input/output allocations. This finding provides a starting point for future empirical studies that focus on the merging behavior of firms. In particular, we outline the analysis of the degree distribution of merging partners in the Japanese Business Firm Network as an important topic for future research. The second elaborated hypothesis of this study states that companies tend to substitute business relations in favor of companies that share many common business partners. This can be deduced from the observed improvement of the model to better reproduce the appearance of motifs when substitution possibilities are taken into account. Providing empirical support for this hypothesis is another subject of future studies.

Despite the specific application to the Japanese Business Firm Network, our results are also potentially relevant for complex networks in other disciplines. For example, in biological networks with high abundances of fully connected motifs with unidirectional links [5], our investigated mechanisms might offer a starting point for future research on the principles underlying the formation of such networks.

In summary, this study provides initial insights motivating future research on influencing motif distributions in network models with a scale-free topology. We consider further systematic studies on correlations between the motif appearances of different patterns as relevant research opportunities. In particular, an assessment of correlations in the course of modifications to the creation and merging processes of nodes would provide further insights into the ability to influence motif appearances on scale-free network topologies.

Acknowledgments

We are grateful towards the Japan Society for the Promotion of Science (JSPS) for financially supporting J.M. as an International Research Fellow of JSPS. This work was partially supported by a Grant-in-Aid for Scientific Research (B), Grant Number 26310207, HPCI System Research Project (Project ID: hp160261) and by JST, Strategic International Collaborative Research Program (SICORP) on the topic of "ICT for a Resilient Society" by Japan and Israel. We would like to express our appreciation to the Center for Teikoku Databank Advanced Data Analysis and Modeling, Tokyo Institute of Technology for providing the datasets. Furthermore, we acknowledge the European Regional Development Fund (ERDF), the German Federal Ministry of Education and Research and the Land Brandenburg for supporting this project by providing resources on the high performance computer system at the Potsdam Institute for Climate Impact Research. We also thank the German Federal Ministry of Education and Research (BMBF) for supporting this project via the Young Investigators Group CoSy-CC^2 (BMBF grant no. 01LN1306A). Special thanks are attributed to Kotaro Tamura, Hayato Goto, Yuta Murakami and Takumi Sueshige for the fruitful scientific discussions and for providing technical assistance. Finally, we are grateful to the reviewers for their comments helping to prepare the presentation of this manuscript.

References

[1] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwanga D U 2006 *Phys. Rep.* **424** 175–308

[2] Newman M E J 2003 *Siam Rev.* **45(2)** 167–256
[3] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U 2002 *Science* **298(5594)** 824–827
[4] Kashtan N, Itzkovitz S, Milo R and Alon U 2004 *Phys. Rev. E* **70(3)** 031909
[5] Alon U 2007 *Nat. Rev. Genet.* **8(6)** 450–461
[6] Paulau P V, Feenders C and Blasius B 2015 *Scientific Reports* **5** 11926
[7] Zhang X, Shao S, Stanley H E and Havlin S 2014 *Europhysics Letters* **108(5)** 58001
[8] Milo R, Kashtan N, Itzkovitz S, Newman M E J and Alon U 2003 *arXiv:cond-mat/0312028*
[9] Xu X K, Zhang J and Small M 2011 *Physica A* **390(23-24)** 4621–4626
[10] Angulo M T, Liu Y Y and Slotine J J 2015 *Nat. Phys.* **11(10)** 848–852
[11] Albert R and Barabási A L 2002 *Rev. Mod. Phys.* **74** 47
[12] Goto H, Takayasu H and Takayasu M 2015 *Proceedings of the International Conference on Social Modeling and Simulation, plus Econophysics Colloquium 2014*
[13] Serrano A and Boguna M 2003 *Phys. Rev. E* **68** 015101(R)
[14] He J and Deem M W 2010 *Phys. Rev. Lett.* **105** 198701
[15] Fagiolo G, Reyes J and Schiavo S 2009 *Phys. Rev. E* **79** 036115
[16] Maluck J and Donner R V 2015 *PLOS ONE* **10(7)** e0133310
[17] Ohnishi T, Takayasu H and Takayasu M 2010 *J. Econ. Interact. Coord.* **5** 171–180 ISSN 1860-711X, 1860-7128
[18] Fujiwara Y and Aoyama H 2010 *Eur. Phys. J. B* **77(4)** 565–580
[19] Takayasu M, Sameshima S, Watanabe H, Ohnishi T, Iyatomi H, Iino T, Kobayashi Y, Kamehama K, Ikeda Y, Takayasu H and Watanabe K 2007 *Ann. Rep. Earth Simul. Cent.* 263
[20] Newman M E J 2001 *Phys. Rev. E* **64(1)** 016132
[21] Amaral L A N, Scala A, Barthelemy M and Stanley H E 2000 *Proc. Natl. Acad. Sci.* **97** 11149–11152
[22] Redner S 1998 *Eur. Phys. J. B* **4(2)** 131–134
[23] Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H E and Makse H A 2010 *Nat. Phys.* **6** 888893
[24] Hein O, Schwind M and König W 2006 *Wirtschaftsinformatik* **48(4)** 267–275
[25] Jeong H, Tombor B, Albert R, Oltvai Z N and Barabási A L 2000 *Nature* **407** 651–654
[26] Broder A, Kumar R, Maghoul F, Raghavan P, Rajagopalan S, Stata R, Tomkins A and Wiener J 2000 *Comp. Net.* **33** 309
[27] Pastor-Satorras R, Vázquez A and Vespignani A 2001 *Phys. Rev. Lett.* **87** 258701.
[28] Ravasz E and Barabási A L 2003 *Phys. Rev. E* **67(2)** 026112
[29] Dorogovtsev S N and Mendes J F F 2002 *Adv. Phys.* **51** 1079
[30] Miura W, Takayasu H and Takayasu M 2012 *Phys. Rev. Lett.* **108(16)** 168701
[31] Fagiolo G 2007 *Phys. Rev. E* **76** 026107
[32] Garlaschelli D and Loffredo M I 2004 *Physical Review Letters* **93(26)** 268701
[33] Leyvraz F 2003 *Physics Reports* **383(2-3)** 95–212
[34] Newman M E J 2005 *Contemporary Physics* **46(5)** 323–351
[35] Moeller S B, Schlingemann F and Stulz R M 2004 *Journal of Financial Economics* **73(2)** 201–228
[36] Petmezas, D 2009 *Journal of Multinational Financial Management* **19(1)** 54–74
[37] Miura W, Takayasu H and Takayasu M 2012 *Eur. Phys. J. Special Topics* **212** 65–75