Front. Environ. Sci. Eng. 2017, 11(1): 11
DOI 10.1007/s11783-017-0903-0

RESEARCH ARTICLE

Diffusion of municipal wastewater treatment technologies in China: a collaboration network perspective

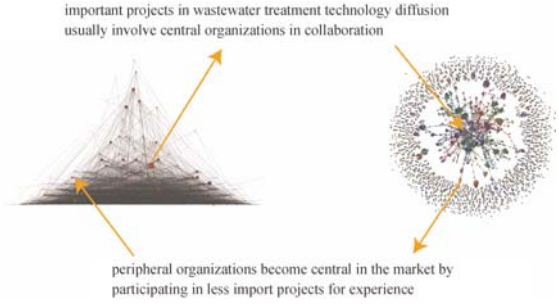
Yang Li[1], Lei Shi (✉)[1], Yi Qian[1], Jie Tang[2]

1 State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing 100084, China
2 Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

HIGHLIGHTS

- Real wastewater treatment technology diffusion process was investigated.
- The research is based on a dataset of 3136 municipal WWTPs and 4634 organizations.
- A new metric was proposed to measure the importance of a project in diffusion.
- Important projects usually involve central organizations in collaboration.
- Organizations become more central by participating in less important projects.

GRAPHIC ABSTRACT



ARTICLE INFO

Article history:
Received 29 August 2016
Received in revised form 26 December 2016
Accepted 26 December 2016

Keywords:
Innovation diffusion
Collaboration network
Wastewater treatment plant
Complex network
Data driven

ABSTRACT

The diffusion of municipal wastewater treatment technology is vital for urban environment in developing countries. China has built more than 3000 municipal wastewater treatment plants in the past three decades, which is a good chance to understand how technologies diffused in reality. We used a data-driven approach to explore the relationship between the diffusion of wastewater treatment technologies and collaborations between organizations. A database of 3136 municipal wastewater treatment plants and 4634 collaborating organizations was built and transformed into networks for analysis. We have found that: 1) the diffusion networks are assortative, and the patterns of diffusion vary across technologies; while the collaboration networks are fragmented, and have an assortativity around zero since the 2000s. 2) Important projects in technology diffusion usually involve central organizations in collaboration networks, but organizations become more central in collaboration by doing circumstantial projects in diffusion. 3) The importance of projects in diffusion can be predicted with a Random Forest model at a good accuracy and precision level. Our findings provide a quantitative understanding of the technology diffusion processes, which could be used for water-relevant policy-making and business decisions.

1 Introduction

Originated in the 18th century and matured in the mid-20th century, modern wastewater treatment technologies have contributed significantly to solve water pollution problems in developed countries [1]. However, many developing countries are still suffering from water pollution in the cities, and the application of municipal wastewater treatment technologies is critical to solve their water-related problems [2]. As the largest developing country, China has built more than 3000 municipal wastewater treatment plants (WWTPs) in the past three decades. With the increase of WWTPs, China's municipal wastewater treatment rate has increased from less than 10% in the 1980s to 90.2% in 2014 [3], which achieved a reduction of 12 million tons of chemical oxygen demand (COD) in 2014 [4] and effectively controlled water pollution problem. Corresponding to the increase of treatment facilities, the diversity of municipal wastewater treatment technology grows quickly in this period. From traditional activated sludge process introduced in the early 1980s to membrane bioreactor in recent years, China has become an

✉ Corresponding author
E-mail: slone@tsinghua.edu.cn

experimental field for over 20 treatment technologies. A thorough analysis of the rapid development and the diversification process in China will contribute to other developing countries on the selection and application of wastewater treatment technologies.

Several articles have concerned about wastewater treatment in China, including the current state [5,6], business models [7,8] and technology evolution [9,10]. Considering the importance of technological diffusion to water pollution control, we mainly explore the patterns and underlying mechanism of the diffusion of municipal wastewater treatment technologies in China.

Based on previous studies, the diffusion of wastewater treatment technology is likely to follow the general pattern of technology diffusion, but also has its own features. Kemp and Volpi [11] have summarized 10 stylized facts for clean technology diffusion, which is consistent with general findings in other fields [12,13]. Some driving factors are identified in almost all technology diffusion cases, such as the economic costs [14], the heterogeneity of users [15], and especially the network of information and interactions [16]. However, as Allan et al. [17] pointed, the key feature that distinguishes green technologies is the generation or facilitation of environmental externalities. The adoption of a wastewater treatment technology in a project may suffer from higher operation cost for achieving the externality. To maintain competitive advantage, organizations involved in these projects need more collaborations to share information, knowledge, resources, etc. [18]. The interaction between demonstrative project and the accumulation of social capital is another worth testing question.

In this article, the relations between the diffusion of wastewater treatment technology and the collaboration of organizations are investigated to answer three questions: 1) can we quantify the importance of projects by their impacts on future technology diffusion? 2) What is the quantitative relation between the importance of projects in diffusion and the importance of organizations in collaboration? 3) Can we make a better decision for technology diffusion with these results?

To answer these questions, we adopted a data-driven complex network approach. Compared with classical diffusion models [19,20] and agent-based models [21,22], data-driven approach can summarize the diffusion patterns directly from fine-grained historical data, which avoids over-simplification and the lack of validation problems. The technology diffusion research based on patent data [23,24] is a typical kind of data-driven research. The research field of complex network emerged in the past two decades with the help of large volumes of data. The diffusion phenomenon is one of the central topics in complex network and machine learning field [16,25], and the researches [26,27] on knowledge diffusion in scientific collaborations provides a new perspective for our interest in technology diffusion in the collaboration of projects.

Following the introduction part, we briefly describe our methodology to collect data, construct network and evaluate importance in Sect. 2. The results section will first describe general properties of the networks in Sect. 3.1, followed by statistical test of the dependence between these networks in Sect. 3.2, and a prediction model using the dependence in Sect. 3.3. The implication and limitation of our work are discussed in Sect. 4 and all results are summarized in Sect. 5.

2 Material and methods

2.1 Data collection

As stated above, our work intends to study the diffusion of municipal wastewater treatment technology in real-world applications, and consider the relationship between collaboration and diffusion. Therefore, we need to collect at least the following information of municipal WWTP projects: 1) their treatment technology, 2) the year when they were built, 3) major participants of each project. Although other parameters (size, location, operation mode, etc.) of the projects may have an influence on the diffusion of technologies, we will mainly consider the participants of each project, which are the "carrier" of technological knowledge. With this information, we can define the diffusion of innovation among projects and the collaboration among organizations.

However, it is not easy to get these required data in China. Although China has experienced a rapid growth in environmental infrastructures, no open data meets our needs directly. We combined data from different data sources and built our own WWTP database. The data sources include government reports and lists on the official websites of the Ministry of Environmental Protection and the Ministry of Housing and Urban-Rural Development, scientific literatures, websites of major organizations that reports its previous projects, field reports, interviews with managers and experts, and a private database from a consulting company. Our final database contains 3136 WWTP projects, which covers the vast majority of WWTP projects from 1981 to 2011, and is sufficient to reveal the diffusion of wastewater treatment technology in China.

Due to the diversity of data sources, one organization might have different names in our raw data. For example, a company name might include the name of a city in one record and without that name in another one; the subsidiary of a company might use a slightly different name from its parent company. Data cleaning is thus a necessary and time-consuming part before analysis; otherwise, we can not create correct connections between these entities. Natural language processing algorithms are employed to distinguish the "core part" of a name, and different names were automatically combined for the same entity. We also

manually checked and corrected the names to ensure the quality of our data. There are 4634 unique organizations left after cleaning, which generally meets our analytical needs.

2.2 Construction of diffusion and collaboration networks

The diffusive and collaborative relations are naturally suitable to be transformed into a network structure. A network G is a mathematical structure composed of a set V of nodes$v \in V$ and a set E of edges$e \in E$. The undirected (or directed) edge can be represented as an unordered (or ordered) pair of nodes, formally $e = \{v_1, v_2\}$ (or $e = (v_1, v_2)$). Besides topological notions, we can also assign attributes to nodes and edges, which is useful to "filter" only part of the network for a specific analysis. In this article, we will define two kinds of networks: one is the diffusion of technology between WWTP projects; the other is the collaboration between different organizations.

To describe the diffusion of knowledge and experience among projects, each project is a node in the diffusion network, and each node has the attributes of its treatment technology and built year. If two projects 1) adopts the same treatment technology (e.g., they both use Oxidation Ditch), 2) share at least one organization (e.g., they were designed by the same institute), and 3) two projects were built in different years (e.g., project A in 2000 and project B in 2002), we assume the knowledge and experience passed from the earlier project to the later one, and add an direct edge pointing from the earlier node to the later one (i.e., $e = (A, B)$). The whole diffusion network from 1981 to 2011 includes 3136 nodes and 8712 directed edges.

The nodes in collaboration network are different organizations. Two organizations will have an undirected edge between them if they have collaborated in the same WWTP project (e.g., organization a and organization b were in the same project, $e = \{a, b\}$), and each edge has three attributes: name of the project, year of collaboration (i.e., built year of the project) and the corresponding treatment technology. The whole collaboration network from 1981 to 2011 includes 4634 nodes and 7204 undirected edges. A visualization of the aggregated diffusion network and collaboration network of all years are shown in Fig. 1.

Given these network representation, we can calculate some network-level properties to understand system evolution, which include the size of giant components, assortativity, etc. The giant component is the largest set of nodes, in which each node can be reached from any other node within. In Fig. 1 (b), the giant component is the main cluster in the center of the graph. Assortativity is a property to measure how similar node would connect with each other. Newman [28] has discussed this property and defined the assortativity coefficient for degree as the Pearson correlation coefficient of pairs of degrees.

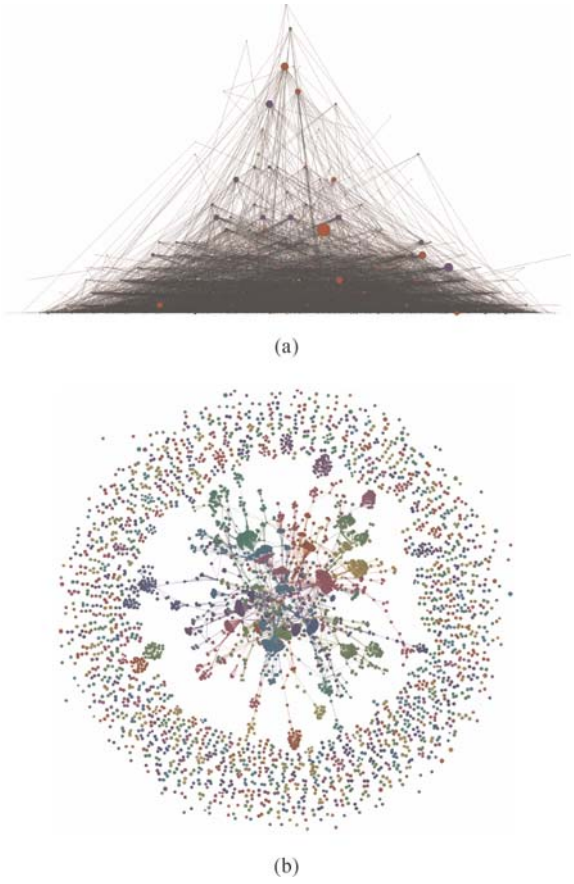


Fig. 1 Visualization of the aggregated (a) diffusion network (layered layout by the built year) and (b) collaboration network (force directed layout)

2.3 Measure the importance of projects and organizations in networks

The constructed networks now enable us to apply metrics and algorithms in network science and understand system characteristics. We will briefly introduce the meaning of major metrics used in our analysis, and an introductory textbook of network science [29] may be helpful if further mathematical and algorithmic details are needed.

One of our key research questions is how to evaluate the importance of a WWTP project in the diffusion of a specific treatment technology. First, it is natural to believe that earlier projects should be generally more important than the later ones, so we should consider the chronological factor (i.e. the direction of diffusion) in this metric. However, if an early project didn't have an impact on others, it is unreasonable to assert its experience and knowledge has contributed to the technology diffusion. An important project should be able to spread its experience to many other projects and make the followers important projects. Therefore, we define the importance of a project as an aggregation of its descendants' importance. Last, despite the importance of direct relations between projects,

innovation might be transferred via other pathways not in the network (such as the job-hopping of experienced engineers or cooperation with foreign experts), and we need to include this factor into our metric. If we use $A_{ij} = 1$ or 0 to represent the existence of directed edge from project i to j, matrix $\boldsymbol{A}$ is the adjacency matrix of the diffusion network. The importance of a project in the diffusion is defined as:

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{\sum_k A_{kj}} + \frac{1-\alpha}{N}. \quad (1)$$

In this equation, x_i is the importance of project i and has two parts linearly combined by a weight parameter α. The first part is the sum of importance of its "descendants" divided by the number of their "ancestors", which considers both the diffusion direction and the number and significance of descendants. The second part means all N projects have an equal importance $1/N$ from other pathways, which is a reasonable assumption without further information. This metric of importance has the same mathematical form as the famous PageRank metric used by Google [30], where the weight parameter usually uses the value 0.85. PageRank value can be iteratively computed via the above equation, or be analytically solved as the eigenvector of a reformed matrix.

Another problem is how to identify dominating organizations in the collaboration networks. We chose three frequently used centrality metrics in diffusion research to measure these organizations. The most frequently used metric for a node is "degree", which was mentioned earlier. In our network model, the larger the node's degree is, the more collaborators it has, and the more popular the corresponding organization is in the market. This metric is a good indicator of an organization's direct influence and reveals local information of a node. However, a node with a high degree value but isolated from the majority is not necessarily significant in collaboration. To take collaborations beyond direct neighbors into account, we also use "k-core" value of a node, which means the node belongs to a maximal subgraph in which each vertex has at least degree k. A node with high "k-core" value has a dense group of collaborators, which is an effective indicator in the spreading of information [31]. "Betweenness" is another frequently used centrality metric. We can define a "path" between two nodes as an alternating sequence of connected nodes and edges, and a "shortest path" as the path with the least number of nodes and edges. According to Freeman [32], this metric is defined as $c_B(v) = \sum_{s \neq t \neq v \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$, where $\sigma(s,t|v)$ is the number of shortest paths between s and t that include node v, and $\sigma(s,t)$ is the number of all shortest paths between s and t. If a node has a large betweenness value, it bridges many different nodes and has a great impact on the flow of information in the system.

Given the importance of projects and the centrality of organizations, we can use statistical methods to summarize the relations between these variables and make predictions based on these relations. In China, organizations with many collaborators are considered to have a good "Guanxi" (relation) with others, and this is believed to be a key to success. On the other hand, the participation in a project may prove the organizations' ability for a specific technology, and help the organizations accumulate experiences to gain advantages in future markets. Based on the exploratory analysis of structural networks, we will use appropriate statistical tests to check whether central organizations make their projects successful in the diffusion of technology, and whether participating in important projects help organizations become central in the collaborations. According to the characteristics of variables and summarized relations, we can choose suitable data mining algorithms and make predictions for policy demands.

In the next section, we will first observe network-level structural property of the whole network to understand the general evolution of diffusion and collaboration processes. Node-level importance will be further investigated to uncover the co-evolution of two networks.

3 Results

3.1 Structures of diffusion and collaboration networks

3.1.1 Structural properties of the diffusion networks

Since there are new municipal WWTP projects built each year, the diffusion network will gradually get new nodes at its boundary and grow larger. As presented in the previous section, we will use the aggregated network from 1981 to a specific year to evaluate the importance of previous projects. Table 1 summarized basic properties for these diffusion networks.

In the first decade (1981–1990), only a few municipal WWTP projects were built (less than 30 nodes), and few of them share the same organization (only 1 edge). In this period, China is experiencing the first wave of economic development, and most people did not realize the value of environmental protection. In the second decade (1991–2000), the numbers of new projects begin to increase, while the linkages between different projects have a much slower increasing rate (only 35 by the end of 2000). The proportion of the largest connected component in the network dropped to less than 5% by the end of this period, implying a fragment market of distributed organizations and an inefficient situation for innovation diffusion. Another indicator assortativity is not so stable and

Table 1 Properties of diffusion networks in 1981–2011

year	new projects	nodes	edges	proportion of largest connected component	assortativity
1981	1	1	0	1.000	insufficient data
1982	0	1	0	1.000	insufficient data
1983	2	3	0	0.333	insufficient data
1984	3	6	0	0.167	insufficient data
1985	1	7	0	0.143	insufficient data
1986	6	13	0	0.077	insufficient data
1987	2	15	0	0.067	insufficient data
1988	3	18	1	0.111	insufficient data
1989	5	23	1	0.087	insufficient data
1990	4	27	1	0.074	insufficient data
1991	5	32	3	0.094	–0.500
1992	5	37	3	0.081	–0.500
1993	8	45	3	0.067	–0.500
1994	7	52	5	0.058	0.167
1995	7	59	5	0.051	0.167
1996	4	63	5	0.048	0.167
1997	10	73	9	0.041	0.060
1998	18	91	12	0.033	0.158
1999	23	114	20	0.035	–0.275
2000	28	142	35	0.049	0.506
2001	62	204	92	0.078	0.740
2002	89	293	167	0.078	0.447
2003	133	426	319	0.089	0.545
2004	155	581	612	0.093	0.424
2005	144	725	841	0.103	0.496
2006	252	977	1202	0.089	0.521
2007	359	1336	2101	0.093	0.602
2008	356	1692	3134	0.112	0.581
2009	685	2377	4866	0.126	0.517
2010	661	3038	7989	0.138	0.426
2011	98	3136	8712	0.136	0.401

sometimes turned negative, which means the connection pattern between projects is not fixed yet. In the third decade (since 2001), environmental problems gradually become the concern of both the public and the government in China. Vast numbers of municipal WWTP projects were built, especially after the bloom of blue algae in Taihu Lake in 2007. The edges between projects increase quickly and the proportion of largest connected component grows, which implies innovation and knowledge are effectively accumulated and passed into new projects. Assortativity remains above 0.4 in this decade, which suggests that successful projects in diffusion are likely to have successful descendant in diffusion, which is consistent with our importance metric. What's more, the number of projects in 2011 is not consistent with the previous trend, suggesting an incompleteness in data collection.

Another feature worth noting in the diffusion network is the variation between different treatment technologies. We assumed that only projects with the same treatment technology could be connected in the network, which determined different technologies would have separate clusters in the diffusion network. The diffusion networks of five major treatment technologies are extracted and visualized in Fig. 2.

It could be observed that A^2O, traditional activated sludge and oxidation ditch (OD) have been adopted and constantly used by some organizations since the 1980s. After the year 2000, A^2O and OD gradually formed a dense

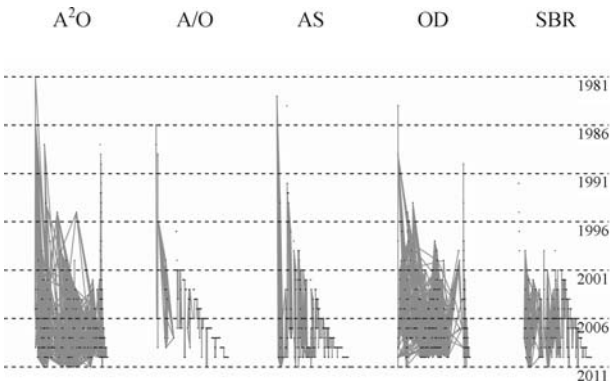


Fig. 2 Visualization of the diffusion networks of five treatment technologies (A^2O = anaerobic-anoxic-oxic, A/O = anaerobic-oxic, AS = traditional activated sludge, OD = oxidation ditch, SBR = sequencing batch reactor)

cluster, and SBR begin to be largely used. However, A/O and traditional activated sludge are not as popular and have less connection between projects, especially after the Taihu Lake event when the removal of nitrogen and phosphorus becomes common demand.

Back to our central question of assessing the importance of project in technology diffusion, in the visualization of diffusion network in Fig. 1, the size of a node is proportional to its importance (PageRank) value by the end of 2011. Earlier projects are generally more important than later ones, and the importance of projects in the same year varies. We are interested in predicting what kind of projects will be an important one compared to others in the same year, which is helpful for policy makers and managers to optimize the resource allocation and promote the fast diffusion of innovation. This question will be discussed in Sect. 3.2, after the analysis of collaborating organizations in the projects.

3.1.2 Structural properties of the collaboration networks

As we can see in Fig. 1, the collaboration network of all years has a "core-periphery" structure: some organizations hold a vital position in the center connected component, surrounded by other small-scale clusters. The largest connected component contains 49% of all organizations involved in WWTP projects in our database. This aggregated network has many interesting properties and unveils the pattern of China's water sector. However, we will mainly focus on the collaboration network in each year, and understand its relation to the importance of projects in technology diffusion.

First, it is common that the collaboration network in a single year is more fragmented than the aggregated one. The size of the largest connected component takes around 20% of all nodes in these networks from the 1990s. We calculated the ratio of the two largest components and plotted this ratio against years in Fig. 3. This ratio is lower than 3 in most of the years, which implies there are matching-level competing groups of organizations in the market. These small groups also exist in the aggregated network as communities, which represent the potential alliance between organizations. There are also some scattered organizations around the big components, which usually serves small projects in Western China and do not participate in the national market. At the end of 2000s, China's water sector started its integration. Dominating organizations cooperated more and more and a giant component emerged in the collaboration network.

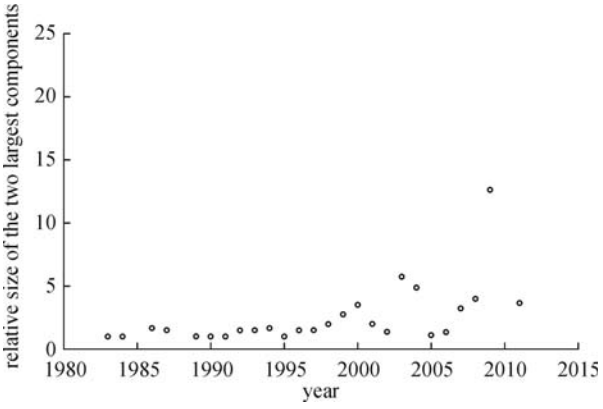


Fig. 3 Evolution of relative size for the two largest components in the network

Another interesting property of the collaboration network is assortativity, which indicates the preference for collaborators. If we have a positive assortativity coefficient for degree, it means the highly centralized organizations tend to connect with each other; while if the value is negative, it means organizations in the center connect more with peripheral ones. Social networks usually have positive assortativity values, and technical networks usually have negative ones. The assortativity of network in different years is shown in Fig. 4. It is interesting that organizations in the central position tend to collaborate with each other in the early years, but the assortativity value decreases and even becomes negative after 2000, which implies more collaborations between central and peripheral organizations. If we look at the detailed network, we find many new organizations enter water market after 2000, and they choose to cooperate with central organizations such as state-owned design institutes or regional water groups. Central organizations can also benefit from this collaboration for taking more control of the projects and expanding their influence.

In conclusion, central organizations in the collaboration network are those highly recognized and preferred by other organizations to cooperate in the market, which also influence the diffusion of wastewater treatment technology.

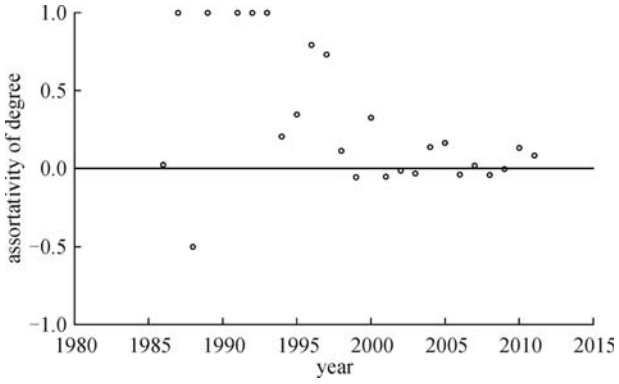


Fig. 4 Assortativity of collaboration networks

3.2 Dependence between diffusion and collaboration network

3.2.1 Effect of collaboration centrality on technology diffusion

Given the diffusion and collaboration networks, we quantified the aspect of collaboration centrality on the importance of a WWTP project in technology diffusion. First, we will investigate if there is a tendency that involving a higher centrality organizations makes a more important project.

According to our measurement of importance, every project will have the same importance in the diffusion network when they were first built. As new projects emerge and inherit their experiences, the importance values will change by year and vary across these projects. The relative importance of all the projects in the same year will converge in 3–4 years. Similar to previous researches [33,34], we assume a project is among the most important ones if 5 years after built it has an importance larger than 80% of all projects built in the same year. We define the set of most important projects as P_{top}, and the set of other projects as P_{ordinary}. The question is then turned into the test of following hypothesis:

Organizations involved in projects from P_{top} are more central than those involved in projects from P_{ordinary} in the collaboration network of their built year.

Since one project usually involves more than one organization, we assume the most central organization in a project dominates the importance in diffusion. The largest centrality value of all organizations in a project is used as the collaboration centrality of the project. The importance of a project is evaluated after 5 years in the corresponding network. We tested three kind of centrality for all projects from 1981 to 2006 against their groups. The kernel density plots in Fig. 5 shows the difference of distributions for P_{top} and P_{ordinary}.

The relationship between collaboration centrality measure and the importance of a project is complicated. There will be no significant relationship if we calculate the linear correlation between them. However, from the distributions exhibited in Fig. 5, it is certain that the most important projects generally have a larger centrality value than other projects. We use the non-parametric Mann–Whitney test to compare the distribution of P_{top} and P_{ordinary} for three kinds of centrality values. The null hypothesis is the distribution of collaboration centrality has no difference between projects in P_{top} and P_{ordinary}, and the alternative hypothesis is the distribution of centrality for projects in P_{top} is stochastically greater than the one for projects in P_{ordinary}.

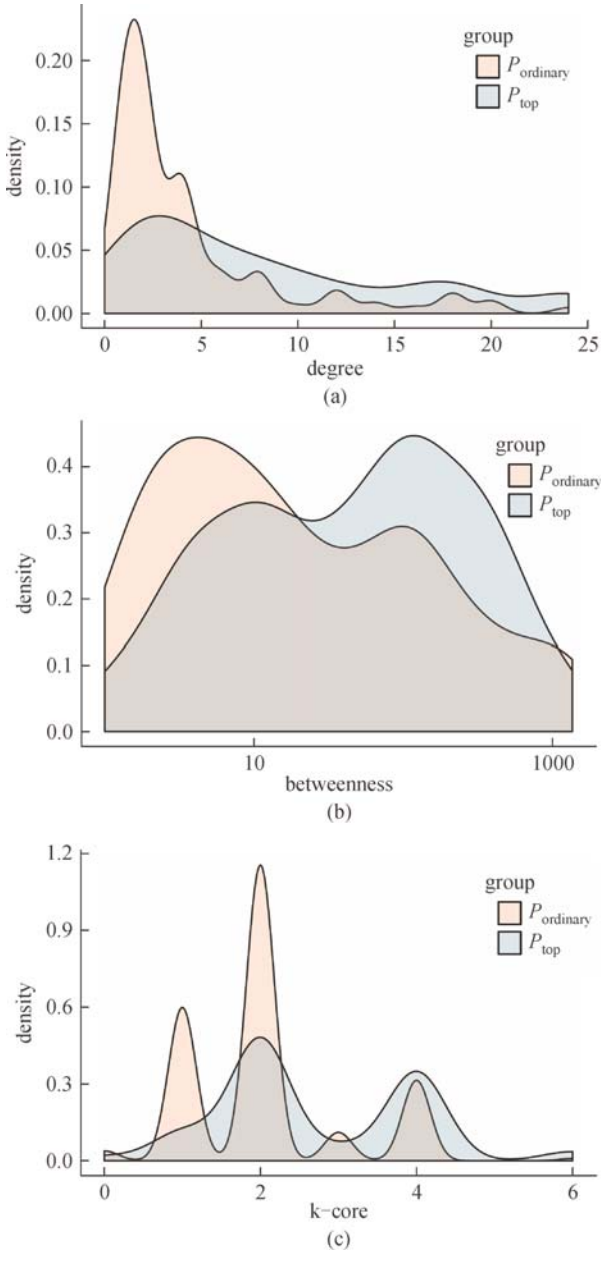


Fig. 5 Kernel density plot of project centrality distributions in P_{top} and P_{ordinary}: (a) kernel density plot of degree; (b) kernel density plot of betweenness; (c) kernel density plot of k-core

The p-values of these one-sided tests and summary statistics are listed in Table S1 in the Supporting Information. All p-values are significantly lower than the 0.05 level. We reject the null hypothesis and believe the distribution of centrality for projects in P_{top} is stochastically greater than the one for projects in $P_{ordinary}$. We also performed the test separately on each year's project and the results are the same, except that in the 1980s the data set is too small for valid testing. The result indicates that central organizations are likely to collaborate more in future projects and spread the knowledge and experience, which makes the project important.

The relationship between the centrality of collaborating organizations and the importance of projects also provide a way to predict successful projects. If a project was done by organizations with a top 20% degree centrality in its built year, the conditional probability for its being among the 20% most important project is 0.524. The conditional probabilities for projects with top 20% betweenness and k-core centralities are 0.469 and 0.448. We will further consider the prediction problem in Sect. 3.3 as a practical application of this result.

3.2.2 Impact of project's importance on future collaboration

In the last section, we found the centrality in collaboration network is a good indicator for important projects in technology diffusion. We have also seen that collaboration network has a preference for central organizations. We will test of following hypothesis to explore the impact of important projects on the centrality of organizations:

Organizations in the collaboration network will be more central after participating in projects from P_{top} than those who participated in $P_{ordinary}$.

As centrality values are growing with time, we calculated the difference of degree, betweenness and k-core centrality values 1 year after and before the organizations participated in a project. Mann–Whitney test was used again and the null hypothesis is the difference for those participated in P_{top} are no larger than those participated in $P_{ordinary}$. The p-values of one-sided test for all centrality and all years approximately equal 1, and we can not reject the null hypothesis. The result means even if an organization participated in an important WWTP project, we can not expect it would become instantly popular in future collaboration. However, the test results are different for an inverse hypothesis:

Organizations in the collaboration network will be more central after participating in projects from $P_{ordinary}$ than those who participated in P_{top}.

The p-values for this inverse hypothesis are smaller than 0.05. We reject the null hypothesis and believe the distribution of centrality increase for organizations in P_{top} is stochastically less than that for organizations in $P_{ordinary}$. The p-value for the test of these two hypothesis are listed in Table S2 in Supporting Information.

We further plotted the difference of an organization's centrality after and before some year against the max importance value of all its projects in the same year in Fig. 6. It is clear that the most probable change in centrality for projects with a high importance is zero, while for less important projects the centrality has a large probability to increase. The result suggests that organizations usually move toward the central position in collaboration networks by participating in less important projects, where they can learn other's experience and accumulate knowledge of a technology. Central organizations may make a project important, but their central position will possibly not change for these projects.

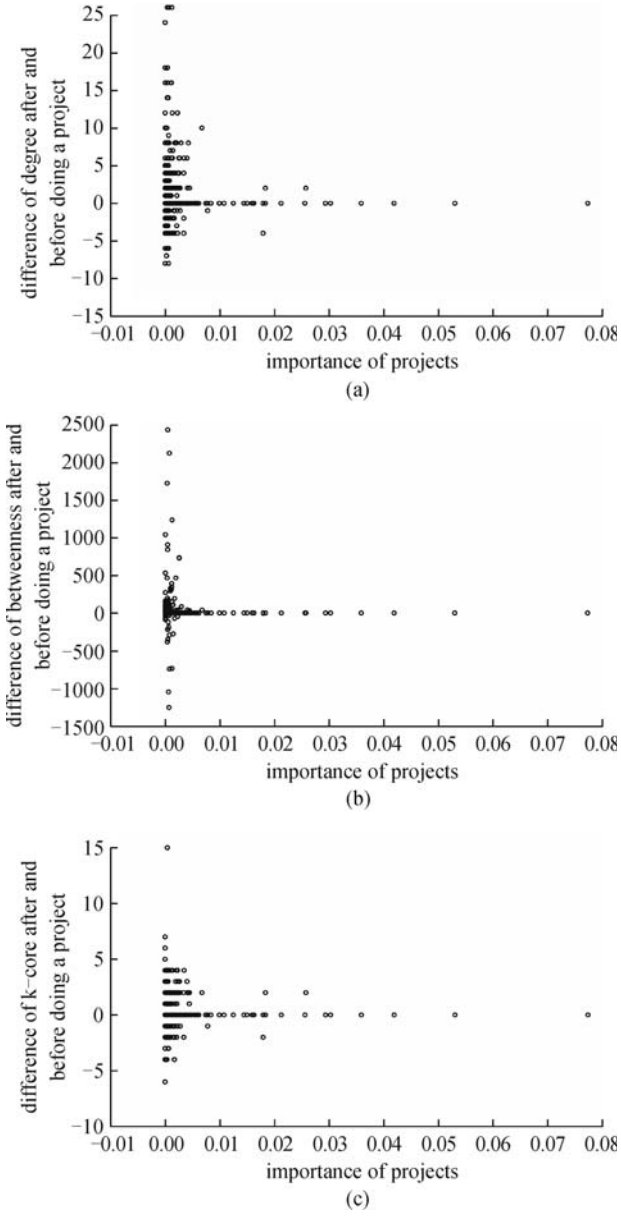


Fig. 6 Difference of an organization's centrality against the max importance value of its projects: (a) the difference of degree; (b) the difference of betweenness; (c) the difference of k-core

3.3 Application: predicting key project for technology diffusion

We have shown in Sect. 3.2.1 that the centrality of organizations in collaboration networks can be used as an indicator for important projects in technology diffusion. However, we can not fully distinguish a project in P_{top} by a single centrality measure. In fact, different centrality measures capture different features of an organization's position in the collaboration, and the relative importance of features may shift across the time. Machine learning models are used in this section to combine weak signals from centrality measures in the collaboration network, and predict whether a project will be among the top 20% most important ones in the diffusion.

Due to the irregular distribution of each centrality measure and the strong correlation between them, we will not use traditional linear methods such as logistic regression here. Instead, we use a Random Forest model [35], which is a combination of decision trees and generally provide accurate predictions for complex data sets. To predict whether a project belongs to P_{top}, we construct a feature vector with the max degree, betweenness and k-core centrality of involved organizations. The type of treatment technology was also included as a feature because of its influence on diffusion. The Random Forest model for one year was trained with all previous years' data, and tested on real outcome of that year to evaluate its performance. We applied this procedure to projects from 2001 to 2006, when there are enough data to build a model and the P_{top} classification to test our result. The sizes of training and testing data set and the performance are summarized in Table 2.

For each project in the training set, the Random Forest model predicts a value for its being in the P_{top}. If the value were larger than a threshold, we would predict the project is among the most important ones. Receiver operating characteristic (ROC) curve is the false-positive rate (ratio that predicted important project are actually not important) against the true positive rate (ratio that predicted important project are indeed important). AUC (Area Under the ROC Curve) is a synthetic metric that captures the performance of predicted result with all possible threshold, which is can be seen as the advantage of a model to a random guess. A perfect prediction has an AUC value 1, and most of our models have an AUC value between 0.6 and 0.8. Random Forest models also provide a measure for variable importance, which suggests that the type of technology is the most effective in prediction, followed by the betweenness, degree and k-core centralities.

For a specific purpose, we can decide our tolerance for errors in different groups and choose a threshold for separation. Our models have an overall accuracy above 0.7 for most range of the threshold, and have an acceptable precision ranging 0.3–0.5 for predicting important projects, compared to the precision of 0.2 for random guess. Although we have seen the overall statistical dependence on centrality, it is not easy to predict the importance of a specific project, which may also be affected by its size, location, operation mode, etc. The predictive model could be further improved if we collected more data and did some feature engineering to create a more complex model, which however demands high quality data and exceeds the scope of this article.

4 Discussion

We will discuss some implications and limitations of our analysis for water innovations in this section.

First, the diffusion of technology via WWTP projects and the collaboration between different organizations are co-evolving processes and depend on each other. Organizations in the collaboration network participate in a project and accumulate knowledge of a treatment technology. When they collaborate with other organizations in other projects, their experiences bridge different projects and spread to more organizations. The diffusion networks of projects and collaboration networks serve as a path of practical information and hidden knowledge, which is critical for the spreading of technology.

Second, important projects in diffusion can be identified and predicted, which may help future decisions. The importance measure proposed in this article reveals the "influence" of a project in the diffusion process, and enables us to identify critical projects from data. The statistical test and machine-learning model further provide a method to predict the important projects with the centrality of organizations in collaboration networks. When the government wishes to promote the diffusion of

Table 2 Training and testing set and the performance of Random Forest models

year of projects in training set	number of projects in training set	year of projects in testing set	number of projects in testing set	AUC
1981–2000	142	2001	62	0.733
1981–2001	204	2002	89	0.837
1981–2002	293	2003	133	0.741
1981–2003	426	2004	155	0.639
1981–2004	581	2005	144	0.692
1981–2005	725	2006	252	0.660

Note: AUC = Area Under the ROC (receiver operating characteristic) Curve

a certain kind of treatment technology, it might be a good idea to work with some central organizations in the collaboration networks and build some demonstration projects.

Third, organizations need some strategies to become dominating in the water market, which is a long-term process. Although central organizations probability make a project important in diffusion, participating in these important projects will likely not help organizations become more central in the collaboration instantly. Peripheral organizations in the network can learn from the central ones, and get the recognition of market by applying the knowledge to less important projects and build their own reputation.

Finally yet importantly, the relations between diffusion and collaboration are statistical results found in a data-driven research. Although we have managed to collect and clean the data set as much as possible, it may still suffer from the incompleteness and incorrectness. The statistical relations are not necessarily causal relationships, and might be the result of some other factors like the technical ability and the financial scales. However, it will be a hard job to collect data or measure these factors for over 4000 organizations and over 3000 projects. The analysis from our limited amount of data at least provides an effective measure of the problem, and can serve practical needs if used properly.

5 Conclusions

In this article, we used a data-driven approach to explore the relationship between the diffusion of wastewater treatment technology and organization collaborations. We collected the data of 3136 municipal WWTPs and 4634 collaborating organizations in China from 1981 to 2011, and transformed the data into a network form for analysis. We have found that:

1) The diffusion networks have a positive assortativity, and different technologies have different structures of networks; while the collaboration network is fragmented, and has an assortativity around zero since the 2000s;

2) Important projects in diffusion usually involve central organizations in collaboration networks, but organizations become more central by doing less important projects;

3) A Random Forest model using centrality measures and the type of technology can predict important projects in diffusion with an acceptable accuracy and precision.

These results provide a quantitative understanding of the evolution and diffusion of wastewater treatment technologies in China, which could be used for relevant policy-making (such as resource allocation for demonstration projects) or business decisions (such as strategy to accumulate social capital by doing less important projects). In the future, more complete and accurate data should be collected to enhance our analysis result. In addition, empirical case studies and model-based studies should be combined with the data-driven method for mutual corroboration of the diffusion mechanisms.

Acknowledgements The research was conducted with financial support from the Tsinghua University Initiative Scientific Research Program (No. 20121088096) and the Major Science and Technology Program for Water Pollution Control and Treatment (Nos. 2012ZX07203-004 and 2012ZX07301-005). We also thank Mr. Tao FU in Peking University for his help in the data collection and analysis.

Electronic Supplementary Material Supplementary material is available in the online version of this article at http://dx.doi.org/10.1007/s11783-017-0903-0 and is accessible for authorized users.

References

1. Lofrano G, Brown J. Wastewater management through the ages: a history of mankind. Science of the Total Environment, 2010, 408 (22): 5254–5264
2. Dechezleprêtre A, Haščič I, Johnstone N. Invention and International Diffusion of Water Conservation and Availability Technologies: Evidence from Patent Data. Paris: OECD Environment Working Papers, No. 82, OECD Publishing, 2015
3. National Bureau of Statistics of China. Statistical Communiqué of the People's Republic of China On the 2014 National Economic and Social Development. Beijing: National Bureau of Statistics of China, 2015 (in Chinese)
4. Ministry of Housing and Urban-Rural Development of China. Construction and Operation Report of Municipal Wastewater Treatment Facilities in 2014 Fourth Quarter. Beijing: Ministry of Housing and Urban-Rural Development of China, 2015 (in Chinese)
5. Qu J, Fan M. The current state of water quality and technology development for water pollution control in China. Critical Reviews in Environmental Science and Technology, 2010, 40(6): 519–560
6. Jin L, Zhang G, Tian H. Current state of sewage treatment in China. Water Research, 2014, 66(1): 85–98
7. Choi J, Chung J, Lee D. Risk perception analysis: participation in China's water PPP market. International Journal of Project Management, 2010, 28(6): 580–592
8. Jang W, Lee D, Choi J. Identifying the strengths, weaknesses, opportunities and threats to TOT and divestiture business models in China's water market. International Journal of Project Management, 2014, 32(2): 298–314
9. Du J, Fan Y, Qian X. Occurrence and behavior of pharmaceuticals in sewage treatment plants in eastern China. Frontiers of Environmental Science & Engineering, 2015, 9(4): 725–730
10. Sun Y, Lu Y, Wang T, Ma H, He G. Pattern of patent-based environmental technology innovation in China. Technological Forecasting and Social Change, 2008, 75(7): 1032–1042
11. Kemp R, Volpi M. The diffusion of clean technologies: a review with suggestions for future diffusion analysis. Journal of Cleaner Production, 2008, 16(1): S14–S21
12. Peres R, Muller E, Mahajan V. Innovation diffusion and new product growth models: a critical review and research directions. International Journal of Research in Marketing, 2010, 27(2): 91–106
13. Stoneman P, Battisti G. The diffusion of new technology. In: Hall B,

Rosenberg N, eds. Handbook of the Economics of Innovation. Vol. 2, 1st ed. Amsterdam: North-Holland, 2010, 733–760

14. Hammar H, Löfgren Å. Explaining adoption of end of pipe solutions and clean technologies – determinants of firms' investments for reducing emissions to air in four sectors in Sweden. Energy Policy, 2010, 38(7): 3644–3651
15. Galán J, Olmo R, López-Paredes A. Diffusion of domestic water conservation technologies in an ABM-GIS integrated model. In: Corchado E, Abraham A, Pedrycz W, eds. Hybrid Artificial Intelligence Systems. Berlin, Heidelberg: Springer, 2008, 567–574
16. Delre S A, Jager W, Bijmolt T H A, Janssen M A. Will it spread or not? The effects of social influences and network topology on innovation diffusion. Journal of Product Innovation Management, 2010, 27(2): 267–282
17. Allan C, Jaffe A B, Sin I. Diffusion of green technology: a survey. International Review of Environmental and Resource Economics, 2014, 7(1): 1–33
18. Andersen M M, Foxon T J. The greening of innovation systems for eco-innovation–Towards an Evolutionary Climate Mitigation Policy. In: Proceedings of the Druid Summer Conference 2009. Denmark: DRUID Society, 2009
19. Bass F M. A new product growth for model consumer durables. Management Science, 1969, 15(5): 215–227
20. Chu J, Wang H, Wang C. Exploring price effects on the residential water conservation technology diffusion process: a case study of Tianjin city. Frontiers of Environmental Science & Engineering, 2013, 7(5): 688-698
21. Kiesling E, Günther M, Stummer C, Wakolbinger L M. Agent-based simulation of innovation diffusion: a review. Central European Journal of Operations Research, 2012, 20(2): 183–230
22. Schwarz N, Ernst A. Agent-based modeling of the diffusion of environmental innovations — An empirical approach. Technological Forecasting and Social Change, 2009, 76(4): 497–511
23. Goetzke F, Rave T, Triebswetter U. Diffusion of environmental technologies: a patent citation analysis of glass melting and glass burners. Environmental Economics and Policy Studies, 2012, 14(2): 189–217
24. Hall B H, Helmers C. Innovation and diffusion of clean/green technology: can patent commons help? Journal of Environmental Economics and Management, 2013, 66(1): 33–51
25. Gomez-Rodriguez M, Leskovec J, Krause A. Inferring networks of diffusion and influence. ACM Transactions on Knowledge Discovery from Data, 2012, 5(4): 1–37 (TKDD)
26. Brunson J C, Fassino S, McInnes A, Narayan M, Richardson B, Franck C, Ion P, Laubenbacher R. Evolutionary events in a mathematical sciences research collaboration network. Scientometrics, 2014, 99(3): 973–998
27. Kim J, Perez C. Co-authorship network analysis in industrial ecology research community. Journal of Industrial Ecology, 2015, 19(2): 222–235
28. Newman M E. Mixing patterns in networks. Physical Review E: Statistical, Nonlinear, and Soft Matter Physics, 2003, 67(2): 026126
29. Newman M. Networks: An Introduction. Oxford: Oxford University Press, 2010
30. Page L, Brin S, Motwani R, Winograd T. The Pagerank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford: Stanford InfoLab, 1999
31. Kitsak M, Gallos L K, Havlin S, Liljeros F, Muchnik L, Stanley H E, Makse H A. Identification of influential spreaders in complex networks. Nature Physics, 2010, 6(11): 888–893
32. Freeman L C. A set of measures of centrality based on betweenness. Sociometry, 1977, 40(1): 35–41
33. Sarig L E, Pfitzner R, Scholtes I, Garas A, Schweitzer F. Predicting scientific success based on coauthorship networks. EPJ Data Science, 2014, 3(1): 1–16
34. Newman M. The first-mover advantage in scientific publication. EPL, 2009, 86(6): 68001
35. Breiman L. Random forests. Machine Learning, 2001, 45(1): 5–32