Collaborative knowledge exchange promotes innovation

Tomoya Mori\textsuperscript{a,b,1}, Jonathan Newton\textsuperscript{a}, and Shosei Sakaguchi\textsuperscript{c}

\textsuperscript{a}Institute of Economic Research, Kyoto University, Yoshida-Honmachi, Sakyo-Ku, Kyoto 606-8501, Japan.; \textsuperscript{b}Research Institute of Economic Affairs, Trade and Industry, 11th floor, Annex, Ministry of Economy, Trade and Industry 1-3-1, Kasumigaseki Chiyoda-ku, Tokyo 100-8901, Japan.; \textsuperscript{c}Faculty of Economics, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.

This manuscript was compiled on November 7, 2022

Considering collaborative patent development, we provide micro-level evidence for innovation through exchanges of differentiated knowledge. Knowledge embodied in a patent is proxied by word pairs appearing in its abstract, while novelty is measured by the frequency with which these word pairs have appeared in past patents. Inventors are assumed to possess the knowledge associated with patents in which they have previously collaborated. We find that collaboration by inventors with more mutually differentiated knowledge sets is likely to result in patents with higher novelty.

\section*{Results}

\textbf{W} refers to the set of distinct word pairs in patent \( p \)'s abstract. (Hereafter, a bold capital letter expresses a set, and the corresponding italic letter its cardinality.) The \textit{novelty} of a word pair at a given point in time is measured by the likelihood of its appearance in patents filed in the past \( 13 \). Specifically, the novelty \( n_{wp} \) of word pair \( w \) at time \( t \) is the ratio of (i) the sum of \( W_p = |W_p| \) over all patents \( p \) filed at dates up to and including \( t \), to (ii) the number of these patents that include word pair \( w \). We measure patent novelty by the average novelty of word pairs in its abstract, \( \frac{1}{W_p} \sum_{w \in W_p} n_{wp} \), where \( t_p \) is the patent’s filing time.

We consider collaborative aspects of patent development by focusing on the productivity per inventor pair, following \( \text{(11)} \). \( H_p \) is the set of all inventors who participated in patent \( p \), while \( M_p \equiv \{(i, j) : i, j \in H_p, i \neq j\} \) is the set of pairs of such inventors. The \textit{average pairwise-contribution to the patent’s novelty} is given by

\[
n_p = \frac{1}{M_p} \sum_{w \in W_p} n_{wp}.
\]

Denoting by \( G_{jt} \) the set of patents inventor \( i \) participated in at time \( t \), define \( i \)'s knowledge at \( t \) by \( K_{it} = \cup_{s < t} \cup_{p \in G_{st}} W_p \) and its novelty by \( k_{it} = \sum_{w \in K_{it}} n_{wt} \).

Inventor pair \( \{i, j\} \) has total knowledge \( K_{ijt} = K_{it} \cup K_{jt} \), with novelty \( k_{ijt} = \sum_{w \in K_{ijt}} n_{wt} \), and inventor \( i \)'s differentiated knowledge relative to \( j \) is \( K_{ijt}^D = K_{it} \setminus K_{jt} \), with novelty \( k_{ijt}^D = \sum_{w \in K_{ijt}^D} n_{wt} \).

Knowledge differentiation between \( \{i, j\} \) is evaluated by the geometric mean of their respective differentiated-knowledge shares in the union of their knowledge,

\[
s_{ijt} = \sqrt[k_{ijt}^D]{k_{ijt}^D/k_{ijt}} \in [0, 0.5].
\]

Their average in patent \( p \),

\[
s_p = \frac{1}{M_p} \sum_{(i, j) \in M_p} s_{ijt},
\]

measures knowledge differentiation in \( p \). We focus on patents with \( s_p > 0 \), since \( s_p = 0 \) implies no knowledge exchange as inventors can be indexed so that \( K_1 \subseteq K_2 \cdots \subseteq K_{M_p} \).

We estimate the effect of \( s_p \) on \( n_p \) by the model:

\[
n_p = \beta_0 + \beta_1 s_p + \cdots + \beta_m s_p^m + \gamma \bar{K}_p + \delta M_p + \phi_p + \varphi_p + \tau_p + \varepsilon_p,
\]

controlling for average knowledge size \( \bar{K}_p \), inventor-pair count \( M_p \), (reflecting the costs/benefits of coordination and task specialization), and fixed effects, \( \phi_p, \varphi_p, \) and \( \tau_p \), for firms, classes of International Patent Classification (IPC), and years, respectively. \( \varepsilon_p \) is a stochastic error.

The estimated conditional expectation and quantiles of \( n_p \) indicate a positive association between \( s_p \) and \( n_p \), except for a range of small \( s_p \), while the observed \( s_p \) are spread over the entire feasible range, \( (0, 0.5] \) (Fig. 1).

To see the robustness of the result, we consider citation counts of a patent as an alternative measure of output. Let \( c_p \) be the citation count of patent \( p \) within five years of application, excluding self-citations, where the self-citations include those

\begin{itemize}
  \item of such inventors. The \textit{average pairwise-contribution to the patent’s novelty} is given by
  \[ n_p = \frac{1}{M_p} \sum_{w \in W_p} n_{wp}. \]
  \item Denoting by \( G_{jt} \) the set of patents inventor \( i \) participated in at time \( t \), define \( i \)'s knowledge at \( t \) by \( K_{it} = \cup_{s < t} \cup_{p \in G_{st}} W_p \) and its novelty by \( k_{it} = \sum_{w \in K_{it}} n_{wt} \).
  \item Inventor pair \( \{i, j\} \) has total knowledge \( K_{ijt} = K_{it} \cup K_{jt} \), with novelty \( k_{ijt} = \sum_{w \in K_{ijt}} n_{wt} \), and inventor \( i \)'s differentiated knowledge relative to \( j \) is \( K_{ijt}^D = K_{it} \setminus K_{jt} \), with novelty \( k_{ijt}^D = \sum_{w \in K_{ijt}^D} n_{wt} \).
  \item Knowledge differentiation between \( \{i, j\} \) is evaluated by the geometric mean of their respective differentiated-knowledge shares in the union of their knowledge,
  \[ s_{ijt} = \sqrt[k_{ijt}^D]{k_{ijt}^D/k_{ijt}} \in [0, 0.5]. \]
  \item Their average in patent \( p \),
  \[ s_p = \frac{1}{M_p} \sum_{(i, j) \in M_p} s_{ijt}. \]
  \item measures knowledge differentiation in \( p \).
  \item We focus on patents with \( s_p > 0 \), since \( s_p = 0 \) implies no knowledge exchange as inventors can be indexed so that \( K_1 \subseteq K_2 \cdots \subseteq K_{M_p} \).
  \item We estimate the effect of \( s_p \) on \( n_p \) by the model:
  \[ n_p = \beta_0 + \beta_1 s_p + \cdots + \beta_m s_p^m + \gamma \bar{K}_p + \delta M_p + \phi_p + \varphi_p + \tau_p + \varepsilon_p, \]
  \item controlling for average knowledge size \( \bar{K}_p \), inventor-pair count \( M_p \), (reflecting the costs/benefits of coordination and task specialization), and fixed effects, \( \phi_p, \varphi_p, \) and \( \tau_p \), for firms, classes of International Patent Classification (IPC), and years, respectively. \( \varepsilon_p \) is a stochastic error.
  \item The estimated conditional expectation and quantiles of \( n_p \) indicate a positive association between \( s_p \) and \( n_p \), except for a range of small \( s_p \), while the observed \( s_p \) are spread over the entire feasible range, \( (0, 0.5] \) (Fig. 1).
  \item To see the robustness of the result, we consider citation counts of a patent as an alternative measure of output. Let \( c_p \) be the citation count of patent \( p \) within five years of application, excluding self-citations, where the self-citations include those
\end{itemize}

\section*{Author contributions: Conceptualization: T.M., J.N.; Funding acquisition: T.M.; Theoretical analysis: J.N., T.M.; Empirical analysis: S.S., T.M.; Writing: J.N., T.M.}

The authors declare no conflict of interest.

\footnote{To whom correspondence should be addressed. E-mail: mori@kier.kyoto-u.ac.jp}
Theoretical model and simulation

There is a set of agents, \( I = \{1, 2, \ldots, I\} \), \( I \) even, who collaborate in pairs to invent new knowledge. When agents \( \{i, j\} \) collaborate on a given patent, the novelty \( n \) of the patent is stochastic and follows a known probability distribution \( P(n|s) \)

conditional on the knowledge differentiation \( s \in (0, 0.5) \) between the collaborators.

The value \( v = v(n) \) of an invention is an increasing function of the novelty of the invention. Assume that \( v(\cdot) \ll \infty \) so that \( V(s) \equiv E_{P(\cdot|s)}[v(n)] \ll \infty \).

The cost of collaboration by a pair \( \{i, j\} \) is a continuous, increasing function \( c_{ij}(s) \) of \( s \) between the collaborators. Assume \( c_{ij}(0) > 0 \). Collaboration carries a cost which increases as the knowledge sets of the collaborators become more differentiated, as there is less of a basis for the common understanding that assists effective communication. Various factors, such as geographic proximity and personality traits mean that this cost will be heterogeneous across pairs.

As only profitable collaborations are pursued, the net value of a collaboration between \( \{i, j\} \) is given by \( A_{ij} = \max\{0, V(s) - c_{ij}(s)\} \). For simplicity, assume that \( A_{ij} \) differs across all pairs \( \{i, j\} \) and that when a pair collaborates, \( A_{ij} \) is split equally between members of a pair (see SI for an alternative value split by bargaining). This defines a one-sided non-transferable utility matching problem (14). Under these assumptions, a unique stable Pareto-efficient matching exists.

To find this matching, first match the pair \( \{i, j\} \) with the largest \( A_{ij} \) and remove them from \( I \). Then, repeat this process until all agents match into collaborating pairs.

To simulate a specific instantiation of this model, we consider technological categories with different levels of technological maturity that give rise to different levels of knowledge.
directions. Specifically, assume technological categories \( L \equiv \{1, \ldots, L\} \) with each agent working in a single category. That is, \( I = \Gamma_1 \cup \ldots \cup \Gamma_L \), where \( \Gamma \) represents the set of agents working in category \( l \in L \); \( \Gamma \cap \Gamma^m = \emptyset \) for \( l \neq m \). Assume that if agents \( i \) and \( j \) are in different categories, then \( c_{ij} \) is large enough that such agents never collaborate.

Let \( K_i \equiv \bigcup_{j \in K_i} K_j \) denote the knowledge set specific to category \( l \in L \). Assume each agent’s knowledge set has equal size \( K_i = K(< K_1 \forall l \in L) \) and that all word pairs are equally novel so that \( n_{ij} \) is the same for all \( w \). Note that the maximum feasible value of \( s_{ij} \) is increasing in \( K_i \). In this setup, categories with smaller \( K_i \) can be considered to represent more technologically mature categories since agents have more knowledge in common, hence less room for knowledge recombination in these categories.

Assume that all categories share a value function, \( v(n(s_{ij})) = \tilde{v}(s_{ij})e^{\epsilon_{ij}} \), where \( \tilde{v}(s) \) is given by the quartic function of \( s \) from the estimated conditional expectation of the observed novelty of patents (Fig. 1A), with an appropriately adjusted intercept \( v_0 > 0 \). For agents within the same category, the cost function is given by \( c_{ij}(s_{ij}) \equiv c(s_{ij})e^{\epsilon_{ij}} \) with \( c(s) \equiv c_0 s \) and \( c_0 > 0 \). \( \epsilon_{ij}, \epsilon_{ij} \sim N(0, 1) \) are idiosyncratic noises.

Fig. 3 demonstrates that this model qualitatively replicates the observed variation in \( s \) and \( n \) from inter-category variation in technological maturity and intra-category variation due to mismatch and idiosyncratic costs in collaboration. The mismatch results from failing to achieve the best match due to the finiteness of the set of possible collaborators (see SI).

**Materials and Methods**

**Patent data.** Patent data are taken from the published patent applications of Japan (16). Identical inventors are traced by matching their names and the establishments they belong to. The data includes all patents filed between 1994 and 2017. We focus on those in 2009 and later, 970,197 of which involve multiple inventors. The 1994-2008 data are used to compute the novelty of the patents filed after 2009.

**Word standardization.** We use NLTK python library to standardize words as follows. (i) Nouns and verbs are lemmatized. (ii) Distinction between a verb and a noun is judged in the context. (iii) Numbers, non-alphabetical characters, and single-character words are removed. As an exception, hyphen-connected words (e.g., self-organization) are kept. Noun/verb parts of these words are lemmatized. (iv) Stop-words (e.g., are and also) are removed.

**Data availability.** All data and codes are provided in SI.

**ACKNOWLEDGMENTS.** This research was conducted as part of the project, “Agglomeration-based framework for empirical and policy analyses of regional economies,” undertaken at the RIETI, Japan. This research has also been supported by the Kajima Foundation, and the Murata Science Foundation, the International Joint Research Center of Advanced Economic Theory of the Institute of Economic Research in Japan, and the Grant in Aid for Research Nos. 17H00987, 19K15108, and 21H00695 of the MEXT, Japan.

1. M Tomasello, M Carpenter, Shared intentionality. Dev. science 10, 121–125 (2007).
2. M Tomasello, E Herrmann, Ape and human cognition what’s the difference? Cur. Dir. Psychol. Sci. 19, 3–8 (2010).
3. M Tomasello, H Rakoczy, What makes human cognition unique? from individual to shared to collective intentionality. Mind & Lang. 18, 121–147 (2003).
4. S Wychty, BF Jones, B Uzzi, The increasing dominance of teams in production of knowledge. Science 316, 1036–1039 (2007).
5. J Singh, L Fleming, Lone inventors as sources of breakthroughs: Myth or reality? Manag. Sci. 56, 41–56 (2010).
6. J Singh, D Kryscynski, X Li, R Gopal, Pipes, pools, and filters: How collaboration networks affect innovative performance. Strateg. Manag. J. 37, 1649–1666 (2016).
7. B Uzzi, J Spirs, Collaboration and creativity: The small world problem. Am. journal sociology 111, 447–504 (2005).
8. SD Angus, J Newton, Emergence of shared intentionality is coupled to the advance of cumulative culture. PLOS Comput. Biol. 11, e1004587 (2015).
9. J Newton, Shared intentions: The evolution of collaboration. Games Econ. Behav. 104, 517–534 (2017).
10. H Rusch, The evolution of collaboration in symmetric 2 × 2-games with imperfect recognition of types. Games Econ. Behav. 114 (2019).
Supporting Information

S1. Dataset

Replication package including the datasets, Python and R codes is available from https://www.dropbox.com/sh/ad1qrbkd5nik9x2/AAAAOHuoQho0hSkHoiUruK3ha?dl=0.

S2. Theoretical model and simulation

An alternative matching with bargaining. In the Model section it was assumed that value generated by collaborating agents is evenly split. This simplifies exposition, but no real problem arises if we assume that members of a pair can bargain over the allocation of surplus from a collaboration. In this case the matching problem become a one sided assignment game. There will still be Pareto efficient outcomes (matchings with allocation of surplus). However, these need not be stable. Indeed, the set of stable outcomes may be empty.

Optimal knowledge differentiation in the simulation model. Assume that $K_i = K_j = K$ for all agents, $i, j$, and that $n_w = 1$ for all $w$. This implies $K_{ij}^D = k_{ij}^D = k_{ji}^D = K^D$, so that $s_{ij} = K^D_i / K^D_j$, where $K_{ij}^D = |K_i^D \cup K_j^D|$ and $K_j^D = |K_i \cup K_j|$. Since the feasible values of $K_{ij}^D$ are $0, 1, \ldots, K$, the set of feasible values of $s_{ij}$ is given by

$$s(\tilde{K}) = \left\{ \frac{K^D_i}{K^D_i + K} \middle| K^D_i = 0, 1, \ldots, K \right\}.$$  

Optimal levels of knowledge differentiation between agents $i$ and $j$ for a given size, $K$, of individual knowledge set are given by the set $s_{ij}^*(\tilde{K}) = \arg\max_{s \in s(\tilde{K})} V(s) - c_{ij}(s)$.

Note that, even if $c_{ij}$ is identical across all pairs $\{i, j\}$, in a finite population it may still be impossible for a given agent $i$ to find and match with an agent $j$ satisfying $s_{ij} \in s_{ij}^*(\tilde{K})$. Such an agent may not exist in the population, or may have already been matched to someone else.