Abstract
Given that research in the Global Software Development (GSD) area has increased over the last decade, and is beginning to mature, this paper surveyed hundreds of GSD studies and tried to obtain a consolidated view of the areas witnessing the highest concentration of research and areas receiving scant attention in the Information Systems (IS) discipline. After careful selection, 83 research articles remained in the study sample database due to their desired characteristics: behavioral study in the IS discipline, and publication between the year of 1999 to 2011. We reviewed these articles and conducted an objective data analysis by using Latent Semantic Analysis (LSA) approach. The research results show that there are 5 major research themes in the GSD area: virtue team, knowledge management, control, trust and risk. We also further streamlined these 5 themes to 3C (coordination, control and communication) research stream model. The consolidated view of GSD studies helps identify trends and patterns in this research area. This study also recommends the future research topics in the GSD area.

Keywords
Global Software Development, GSD, Information Systems, coordination, control, communication, virtue team, knowledge management, risk, trust, control, Latent Semantic Analysis
Introduction
Global software development (GSD) relates to software development undertaken at geographically distributed locations across national and international boundaries in a coordinated fashion involving synchronous and asynchronous interaction (Haq, Khan, & Tariq, 2011). GSD involves in communication for information exchange, coordination of distributed teams and their members, and effective control of teams by adhering to goals and policies related to a project. Although GSD is becoming an effective technique, it suffers from many challenges such as poor communication, lack of trust and coordination. These challenges pose serious risk to the smooth execution of the GSD projects. In the past decade, GSD has become an area of active research in many disciplines e.g., Economics, Sociology, Management, Marketing, and Management Information Systems; and a large number of researches have worked on GSD topic from different aspects: for instance, offshore outsourcing advantages -- access to a large pool of qualified human resources, significant cost advantage (Layman, Williams, Damian, & Bures, 2006; Sarker & Sahay, 2004), speeding up of work by “following the sun” (Sarker & Sahay, 2004), and enhanced “strategic flexibility” (Robinson & Kalakota, 2004); and offshore outsourcing challenges -- cultural and language difference, trust and commitment, extended feedback loops, asynchronous communication, and knowledge management.

Many GSD studies have been conducted and provided researchers and practitioners with the lessons learned from GSD projects, and the tactics that can alleviate challenges. For instance, Carmel and Agarwal (2001) provided three tactics: 1) reduce intensive collaboration, 2) reduce cultural distance, and 3) reduce temporal distance. Other strategies were discussed as well such as keeping everyone informed, reducing dissonance or conflict, aiding project performance,
providing sufficient opportunities for face-to-face meetings, and using of more cue-laden communication modes such as video-conferencing (Lee-Kelley & Sankey, 2008).

Given that research in GSD has increased over the last decade, and is beginning to mature, it is perhaps the time to obtain a consolidated view of the areas witnessing the highest concentration of research and areas receiving scant attention evolving the GSD phenomenon. Such consolidated view of a phenomenon, as Dube and Pare (2003, p. 599) suggested, “help identify trends and patterns”, and therefore, “serve as an instrument to reflect as a research community on our progress”. Such a consolidated analysis also enables researchers to understand and pinpoint areas where more work needs to be done (Dube & Pare, 2003). The present study tries to play this role in Information Systems (IS) research discipline on topics in the GSD research stream. This study can contribute in the following ways: first, it focuses on IS-related research studies, especially those with a behavioral focus (versus the technical journals and conference) since this study attempts to be more relevant to IS. Secondly, it studies a large time period (i.e., 1999-2011), thereby enabling the capture of more recent developments. Thirdly, it will identify the critical topics that have been studied. Finally, it will make recommendations to scholars for future research areas in the GSD research stream. Through this examination, the present study attempts to answer the following research questions:

- What topics of GSD research are IS academics currently addressing?
- Which of the identified topics of GSD research are under-explored by IS academic?
- What is the trend in IS academic research in the GSD research?

In order to answer the aforementioned questions, the present study conducted a Latent Semantic Analysis (LSA) on the existing GSD literature. The analysis, based on the 83 GSD articles, reveals that GSD research in IS discipline has 5 major research themes: virtue team, knowledge
management, trust, control, and risk. And these 5 themes can be further classified into three research streams: coordination, which covers virtue team and knowledge management; control, which covers control and risk; and communication, which covers trust. The rest of the manuscript is organized as follows: first, the author describes the research method, followed by the results and discussions section; next, recommendations are presented; finally, it concludes with the contributions and limitations.

**Method**

**Data Collection**
The terms used for searching IS-related articles were “global software development”, “Information Technology outsourcing”, “offshore outsourcing”, and “communication”. The article search was accomplished by using ABI Inform (Proquest) online, Business Source Complete (EBSCO), and Google Scholar. Some articles were searched based on the reference lists of selected articles. The author also searched the premier IS journals such as *MIS Quarterly* and *Information Systems Research (ISR)*.

The search returned hundreds of articles, and the first author read through the abstracts first in order to determine whether these articles were Information Systems (IS) related. After careful selection, 83 articles remained in the database. These 83 articles were used to conduct the data analysis by using the Latent Semantic Analysis approach.

**Data Analysis**

**A Brief Introduction to Latent Semantic Analysis**
Latent Semantic Analysis (LSA) is a well-established text mining technique (Han, Kamber, & Pei, 2011). The technique was originally developed from the area of natural language processing in the computer science discipline. It has been adopted by some popular search engines (i.e.
Google and Bing) as one of their algorithms to process the textual data on the Internet (Hossain, Prybutok, & Evangelopoulos, 2011). In recent years, this technique has been introduced to other disciplines for the purpose of knowledge discovery from scholarly literature. For example, a study published in MIS Quarterly (Sidorova, Evangelopoulos, Valacich, & Ramakrishnan, 2008) adopts this technique to identify the intellectual cores of the entire IS discipline. A recent paper (li & Joshi, 2012) uses LSA to uncover the major research themes for an emerging discipline of social computing.

In contrast to other types of text mining techniques, Latent Semantic Analysis (LSA) is a special mathematical and statistical method used to identify the latent concepts within the textual data at the semantic level (Hossain, et al., 2011). LSA is able to extract the contextual-usage meaning of words and obtain approximate estimates of meaning similarities among words within the given textual data, thus providing the information at the semantic level (Hossain, et al., 2011). LSA has numerous applications in natural language processing, search engine and library indexing and many other areas (Hossain, et al., 2011). LSA simulates the way the human brain distills meaning from text (Sidorova, et al., 2008). The detailed mathematical explanation for LSA can be found in previous studies (Sidorova, et al., 2008). This section focuses on how LSA is conducted in this particular study.

**Operationalization of LSA in This Study**

First of all, we collected the abstracts from the previously identified 83 papers and compiled them into a spreadsheet. Each abstract in the spreadsheet was converted into a document object in Rapidminer 5.0 and each abstract was assigned a unique document ID. Then the documents went through a series of pre-processing procedures. 1) All the letters in these documents were transformed into lower case. 2) The documents then were tokenized with non-letter separators.
As a result, each document was split into a sequence of words (or tokens). 3) We removed the “stopwords” in the identified word list. “Stopwords” include the trivial English words such as “and,” “the,” “is,” “a,” “an” and so on. These stopwords don’t provide meaningful information about the documents and their presence unnecessarily increases the dimensionality. 4) We removed all the tokens that are less than two letters (i.e. “s,” “x,” and so on), because we found those tokens don’t contain meaningful information. 5) We also removed the words or tokens that appear only in one document, because these tokens are associated only with the specific study and shouldn’t be considered as a reflection of any research theme. 6) We applied term stemming techniques to word list. Terms stemming will identify the root of the words and regard all words with the same root as one token. For example, “collaborate,” “collaborating,” “collaboration,” and “collaborative” will be regarded as a single token, the “collabor–.” By doing so, different variants of the same word are combined and the dimensionality is further decreased. 7) Finally, we removed “author,” “paper,” “conclusion” and some other words that are associated solely with the writing style of scholarly articles and don’t provide additional information about the content. All these term reduction steps eventually resulted in a word list with 997 tokens. This word list represents a list of the key words that exist in these 83 collected papers and contain meaningful information which can be used to identify the research themes of the area.

Afterwards, we applied the Term Frequency Matrix Transformation which helps the unique and informative key words be assigned with more weight than common words. Then this matrix was decomposed into three matrices using Singular Value Decomposition. These three matrices are namely the term-by-factor matrix, singular value matrix (square roots of eigenvalues), and the document-by-factor matrix. The term-by-factor matrix shows the term loadings to a particular latent factor. The document-by-factor matrix shows the document loadings to a particular latent
factor. The singular values (square roots of eigenvalues) represent the importance of particular factor. The detailed explanation of Term Frequency Matrix Transformation and Singular Value Decomposition can be found in Li and Joshi’s (2012) paper.

We first examined the singular values in the singular value matrix. Singular value is the square roots of eigenvalues, which theoretically represent the amount of variance explained by a particular factor as well as the importance of a factor. The top 5 most important factors and their singular values are presented in Table 1.

**Factor Interpretation**
The main purpose of the study is to uncover the major research themes of the global software development research; therefore we aimed to select the least number of factors that can cover all the existing content in these 83 papers. To serve this goal, we examined the document-by-factor matrix. In the matrix, each cell shows the loading of a particular paper abstract on a particular factor. The higher loading means the higher degree of relevance between a particular document and a certain latent factor. Negative loading means the document is irrelevant to a particular latent factor. Simply put, the number of positive loadings on a factor indicates the number of papers explained by this factor. We further counted the cumulative number of distinct papers explained by the factors. As shown in Table 1, Factor 1 covers the content of 45 distinct papers. Factor 2, together with Factor 1, cumulatively explain 68 papers. Finally, we found that the top 5 factors collectively explained all 83 papers. Therefore, these five factors represent the major research themes in the GSD area.

**Table 1 Factors and Their Singular Value**

<table>
<thead>
<tr>
<th>Factor</th>
<th>Singular Value (square roots of eigenvalues)</th>
<th>Cumulative number of papers can be explained</th>
<th>High Loading Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor 1</td>
<td>1.474</td>
<td>45</td>
<td>Team, time, member, separ, virtual, collabor, cultur, time_zone, virtual_team</td>
</tr>
</tbody>
</table>
However, these five factors so far were just five symbols; we wouldn’t understand the meaning of them unless the factors were interpreted. We interpreted the meaning of these factors by examining the term-by-factor matrix. As discussed above, terms will load to a particular factor and collectively define the meaning of the factor. The top loading terms of the top five factors are displayed in the Table 1. These terms define the meaning of the top five factors or research themes.

We associated each factor with its high-loading terms and documents to assist factor interpretation. We used these terms and documents to interpret and characterize (i.e., label) the particular factor. The process of labeling the factors consist of examining the terms and abstracts related to a factor, interpreting the underlying area, and determining an appropriate label. Finally, we identified 5 labels, which are also the research themes, in the GSD area. They are virtue team, knowledge management, control, risk, and trust (as shown in table 2).

<table>
<thead>
<tr>
<th>Factors</th>
<th>Labels</th>
<th>High-loading Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor 1</td>
<td>Virtue Team</td>
<td>Team, time, member, separ, virtual, collabor, cultur, time_zone, virtual_team</td>
</tr>
<tr>
<td>Factor 2</td>
<td>Knowledge Management</td>
<td>Knowledge, coordin, product, task, transfer</td>
</tr>
<tr>
<td>Factor 3</td>
<td>Control</td>
<td>Govern, formal, mechan, sla, relat_govern, commit</td>
</tr>
<tr>
<td>Factor 4</td>
<td>Risk</td>
<td>Firm, outsource, client, vendor, contract, cost</td>
</tr>
<tr>
<td>Factor 5</td>
<td>Trust</td>
<td>Practition, cultur_understand, maintain_trust, trust_client, relationship, client_cultur_understand, commun_strategi</td>
</tr>
</tbody>
</table>

Table 2 Interpretation of Factors
In order to find out how these 83 papers are distributed on the five factors, we examined the document-by-factor matrix and counted the number of distinct papers loaded to a particular factor. If one paper has cross loadings to multiple factors, we allocated this paper to the factor on which it has the highest loading. It turns out the virtual team is the biggest research area, which has 23 papers addressing this theme. The knowledge management is the second important research theme, which has 18 papers on it. The control is the third major research theme with 16 papers. The risk and trust themes have 12 and 14 papers on each of them, correspondingly.

![Figure 1 The Web of GSD Research Themes by # of Papers](image)

**Results and Discussions**

**The Core Research Streams of GSD**

The research results show that there are five major themes in the GSD research area: virtue team, knowledge management, control, trust, and risk. As discussed by Sidorove et al. (2008) LSA can be specified to give solutions at different levels of abstraction, which depends on the research
objective to decide what level of abstract is needed. Since the objective of this study is to provide a big picture of the entire GSD study in IS discipline across a long time of period (i.e., 1999-2011), we need the most parsimonious LSA solution to cover the most content. However there is no rule of thumb criterion for doing so (Sidorova, et al., 2008). Therefore, we decided to adopt Carmel and Agarwal’s (2001) impact of distance research framework (Figure 2) in the GSD area to further categorize them into a higher level of abstract.

As many scholars have noticed and discussed the inherited challenge of GSD—the distance issue in their studies (Carmel, 1999; Carmel & Agarwal, 2001; S. Sarker & S. Sahay, 2004). The distance includes not only the geographical distance, but also temporal, and cultural distance. An organization can’t function well without coordination and control; however, distance creates difficulties for coordination and control. Carmel and Agarwal (2001) had thorough discussion of coordination and control, and how they affect each other in the GSD settings. Coordination is the act of integrating each task with each organizational unit, so the unit contributes to the overall objective. Communication is the glue that integrates different units into the organization. Control is the process of adhering to goals, policies, standards, or quality levels. Control can be formal and informal. For today’s knowledge workers, coordination and control have been blended together in many ways. Communication is the mediator that affects both coordination and control. Carmel and Agarwal (2001)’s “Impacts of Distance” figure (Figure 2) shows clear relationships among distance, communication, coordination and control.
From the figure 2, it is seen that distance negatively affects communication, coordination, and control directly; at the same time, distance negatively affects coordination and control indirectly through communication. From this figure, it is clear that effective communication can play an important role in a successful global software development project. Any challenges and issues relate to distance and communication can affect the GSD’s success.

Based on above discussion of distance impact on GSD projects and the identified five factors, we can further classify these five factors into three research streams: coordination, control and communication. Virtue team and knowledge management are highly related to coordination research theme. Control and risk are related to control research theme. And trust is related to communication theme.

**Coordination, Virtue Team and Knowledge Management**

We define coordination here as the management of dependencies among task activities to achieve a goal (Malone & Crowston, 1994). The GSD projects need to be done by at least two teams located dispersal i.e., one is the vendor team and the other one is the client team, which are usually located in different countries. This type of team is called virtue team. The software development task among virtue teams is usually complex and has substantial dependencies that
need to be managed, thus need for coordination. The role of knowledge in software development is important. Software development is considered a knowledge intensive activity of organizing and integrating the specialized expertise, skills, and perspectives of various project stakeholders into an appropriate, coherent, and practical solution (Faraj & Sproult, 2000). The software development in the offshore domain is complex, and the stakeholders need to swiftly move through the learning curves with best practices institutionalized (Rottman & Lacity, 2004). It involves in the knowledge (including explicit-implicit and formal-informal) transfer between client and vendors. This process needs coordination between client and vendor teams intensively.

Control and Risk
Offshore outsourced projects are more likely to be failure than in-house and domestically outsourced because there are greater risks involved. The study shows that half of offshore outsourcing initiatives “fail” or do not meet stated performance objectives because of many risk factors such as the inability to navigate organizational and cultural barriers, middle-management resistance and failure to communicate (Nakatsu & Lacovou, 2009). To avoid the GSD project failure, in other words, in order to alleviate the risk factors in the GSD project, the control mechanism plays an important role here. In the control literature, vendor governance, especially, how the client exercises control over the vendor, has been considered a critical factor to a successful GSD project (Choudhury & Sabherwal, 2003; Clark, Zmud, & McCray, 1998).

Communication and Trust
Communication is fundamental to any form of organizing, especially for the globally distributed teams, for which organizational structure and form can be defined in terms of communication linkages among organizational units (Zack & Mckinney, 1995). These communication linkages are enabled by technology. Advanced ICT enables parties to communicate across distance, time, culture, departments, and organizations, thereby creating an “anyone/anytime/anyplace” form of
organizational experience (Desanctis & Monge, 1999). Effective communication highly relates to the trust among team members (Smith & Blanck, 2002). Javenpaa & Leidner (1999) indicated that lack of trust is likely to develop between globally distributed teams. Smith & Blanck (2002) confirmed this view and found that trust is the foundation for an effective team, but it is also most difficult to build at a distance. Trust is defined as the willingness of one person or group to relate to another in the belief that the other’s action will be beneficial rather than detrimental, even though this cannot be guaranteed (Child, 2001). Trust is more likely to be built if personal contact, frequent interactions, and socializing between teams and individuals are facilitated (Child, 2001). Due to the common barriers of geographical distance, cultural, language, and time difference in the GSD teams, it is unusual that team members can have frequent personal contact, interaction and socialization. Therefore, the frequent formal and informal communication via ICT is important for distributed team members to build trust among them.

<table>
<thead>
<tr>
<th>Research Streams</th>
<th>Factors</th>
<th>Labels</th>
<th>High-loading Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coordination</td>
<td>Factor 1</td>
<td>Virtue Team</td>
<td>Team, time, member, separ, virtual, collabor, cultur, time_zone, virtual_team</td>
</tr>
<tr>
<td></td>
<td>Factor 2</td>
<td>Knowledge Management</td>
<td>Knowledge, coordin, product, task, transfer</td>
</tr>
<tr>
<td>Control</td>
<td>Factor 3</td>
<td>Control</td>
<td>Govern, formal, mechan, sla, relat_govern, commit</td>
</tr>
<tr>
<td></td>
<td>Factor 4</td>
<td>Risk</td>
<td>Firm, outsource, client, vendor, contract, cost,</td>
</tr>
<tr>
<td>Communication</td>
<td>Factor 5</td>
<td>Trust</td>
<td>Practition, cultur_understand, maintain_trust, trust_client, relationship, client_cultur_understand, commun_strategi</td>
</tr>
</tbody>
</table>

Based on the above discussion, we can streamline the five-factor solution to the higher abstractive level of three-stream solution: coordination, control and communication (i.e., 3C research streams model). This 3C model provides us a better view and understanding of the current research trend in the GSD area in IS discipline: all these research studied solutions to alleviate the impact of the distance to the GSD project’s success. They all revolve around three
research streams: coordination, control and communication. The total numbers of papers that studied in each research stream are 41, 28, and 14, correspondingly. These numbers indicate IS scholars paid extremely attention to the coordination issue, and paid least attention to the communication issue in the GSD projects.

**Contributions and Limitations**

**Contributions**
The study can contribute to the IS research field in the following ways: first, it focuses on IS-related GSD research studies, especially those with a behavioral focus (versus the technical journals and conference) since this study attempts to be more relevant to IS; secondly, it studies a long time period (i.e., from 1999 to 2011), thereby enabling the capture of more recent developments in the field; thirdly, it identifies the critical topics that have been studied previously, i.e., virtue team, knowledge management, control, trust and risk; fourthly, it identifies three major research streams in GSD area—coordination, control and communication; fifthly, it reveals the research trend in the GSD area – the most researched stream is coordination, and the least researched stream is communication; therefore, we would like to propose IS scholars to conduct more communication related GSD research studies in the future; finally, this study uses an advanced method, the Latent Semantic Analysis, to conduct a systematic and extensive literature review of publications in the global software development research area. In prior literature review of this kind, researchers’ subjective judgment was used to identify the major research themes, this study makes a novel contribution by using the Latent Semantic Analysis, a mathematical natural language process technique, to synthesize the literature and outline the research landscape of this area.
Limitations
This study has some limitations as well. First, we used ABI Inform (Proquest) online, Business Source Complete (EBSCO), Google Scholar and several other IS journals e.g., MIS Quarterly and Information Systems Research, to compile our sample. Although these are the major databases and journals for the business and IS related publications, it is conceivable that some GSD related research papers were not captured in our sample. Secondly, although LSA is a pure mathematical method that can synthesize large body of literature in the objective manner with very little intervention from the researchers, it only provides certain statistics that requires researchers’ judgment, e.g., labeling each factor. Cautious readers ought to, not only read the labels proposed in this paper, but also review the high loading terms and consider their own interpretation of each factor.

Conclusion
Considering the importance of GSD to modern business, and on a larger scale, to national economies, this research studied a relatively exciting and important area. The present study attempted to explore the development and current situation of GSD research. Through this empirical analysis, we hope to reflect on the “progress” of the body of research within the IS discipline, and therefore brings increased awareness as to the areas requiring further attention.
References:
Han, J., Kamber, M., & Pei, J. (2011). Data Mining: concepts and techniques Morgan Kaufmann.