Data Transformation And Analysis Process for a Social Network Analytics Data Warehouse.

Cite as: Goggins, S. P., Mascaro, C., & Valetto, G. (2013). Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology*, *64*(3), 516-539.

Introduction

This is a fifteen or sixteen step data transformation process. Give or take.

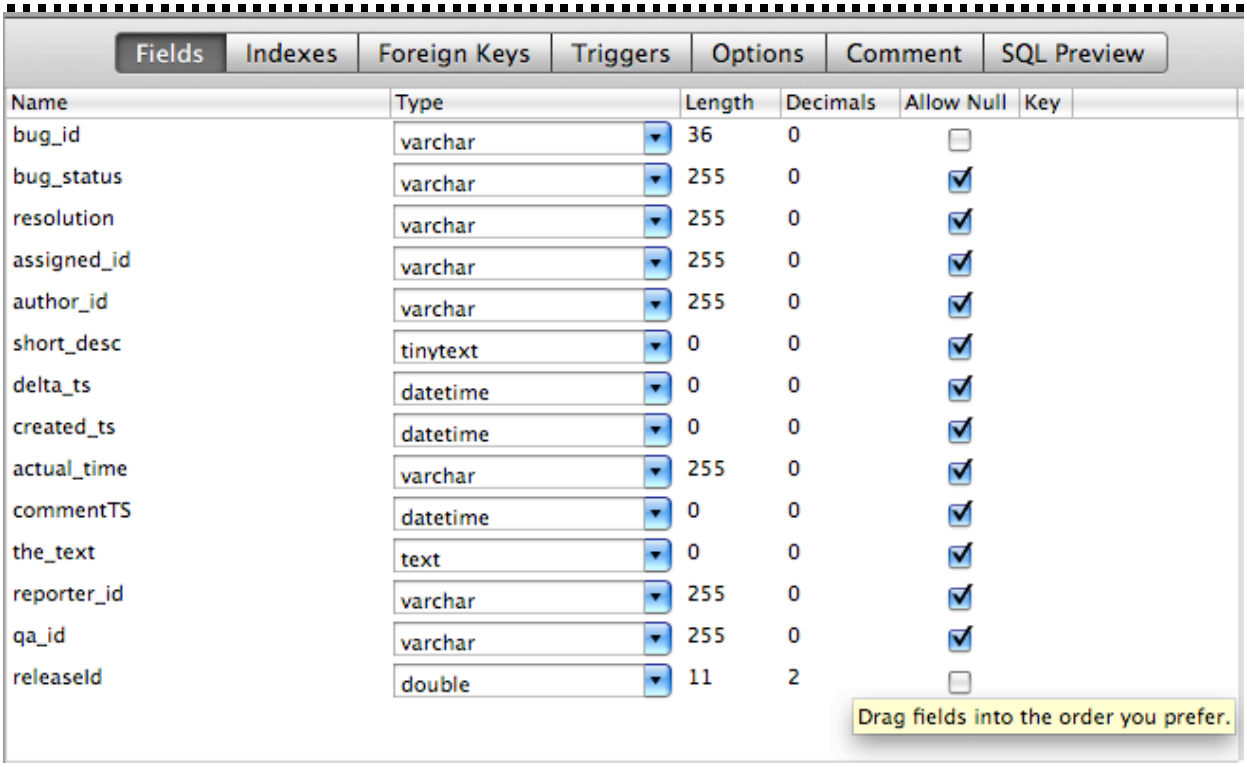
Step 0: Get Data Into the Standard Format

We begin with step 0, which is getting the data from a source system into the required input format: (Fact Table):

Event Fact Column	**Description of Data**
Context	Context is constituted by the context name, the context type and the events_context_id. In the MyLyn data, these elements are "Bugzilla", "Software Engineering" and "MyLyn" respectively. In the CANS data, these elements are "Sakai", "Group" and the course id (there are nearly 100).
Context Type	Software engineering, or "Group" What type of interaction context is this, exactly.
Event Action	The type of action. For now, this is "read" or "create"; jforum.read or jforum.new.
Person – Creator	Person who created the event record
Person – Author	Person who created the object the event record measures
Time	Timestamp of the event, which points to a time dimension, allowing for systematic temporal slicing of the data.
Weight in Minutes	Time distance from the previous event
Row Type	Whether the event row is from source data, or inferred from source data based on approximations of user behavior (i.e., in Sakai, a user is likely to view the five previous posts to some diminishing extent when reading or posting).
Module	Can be mapped to a module in a course, if this data is known.
Semester	Can be mapped based on the dates of the events. An imperfection in the current mapping of mylyn data is that we use the event_session for release metadata. It would be more optimal to leverage the semester as a boundary for release. This is something to explore in future releases. In the future, we may wish to connect institution id and semester in a single version of the fact table.
Institution	Hard coded. Currently the CANS host institution or "Drexel", for the MyLyn data.

Cite as: Goggins, S. P., Mascaro, C., & Valetto, G. (2013). Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology*, *64*(3), 516-539.

Input Data Structure: MyLYN Example



Name	Type	Length	Decimals	Allow Null	Key
bug_id	varchar	36	0	☐	
bug_status	varchar	255	0	☑	
resolution	varchar	255	0	☑	
assigned_id	varchar	255	0	☑	
author_id	varchar	255	0	☑	
short_desc	tinytext	0	0	☑	
delta_ts	datetime	0	0	☑	
created_ts	datetime	0	0	☑	
actual_time	varchar	255	0	☑	
commentTS	datetime	0	0	☑	
the_text	text	0	0	☑	
reporter_id	varchar	255	0	☑	
qa_id	varchar	255	0	☑	
releaseId	double	11	2	☐	

Cite as: Goggins, S. P., Mascaro, C., & Valetto, G. (2013). Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology*, *64*(3), 516-539.

Data Analysis From the Group Informatics LAB Interaction Warehouse

Introduction

Data analysis discussed here begins with data derived from Sean P. Goggins dissertation at the University of Missouri – Columbia. There are rich sets of data, which are described in some detail in two prior publications (Goggins, Laffey, Amelung, & Gallagher, 2010b; Goggins, Galyen, & Laffey, 2010a). We now have a wider array of data in the warehouse, but it is not as diverse methodologically as Goggins dissertation work. The data we have to work with is described in figure one.

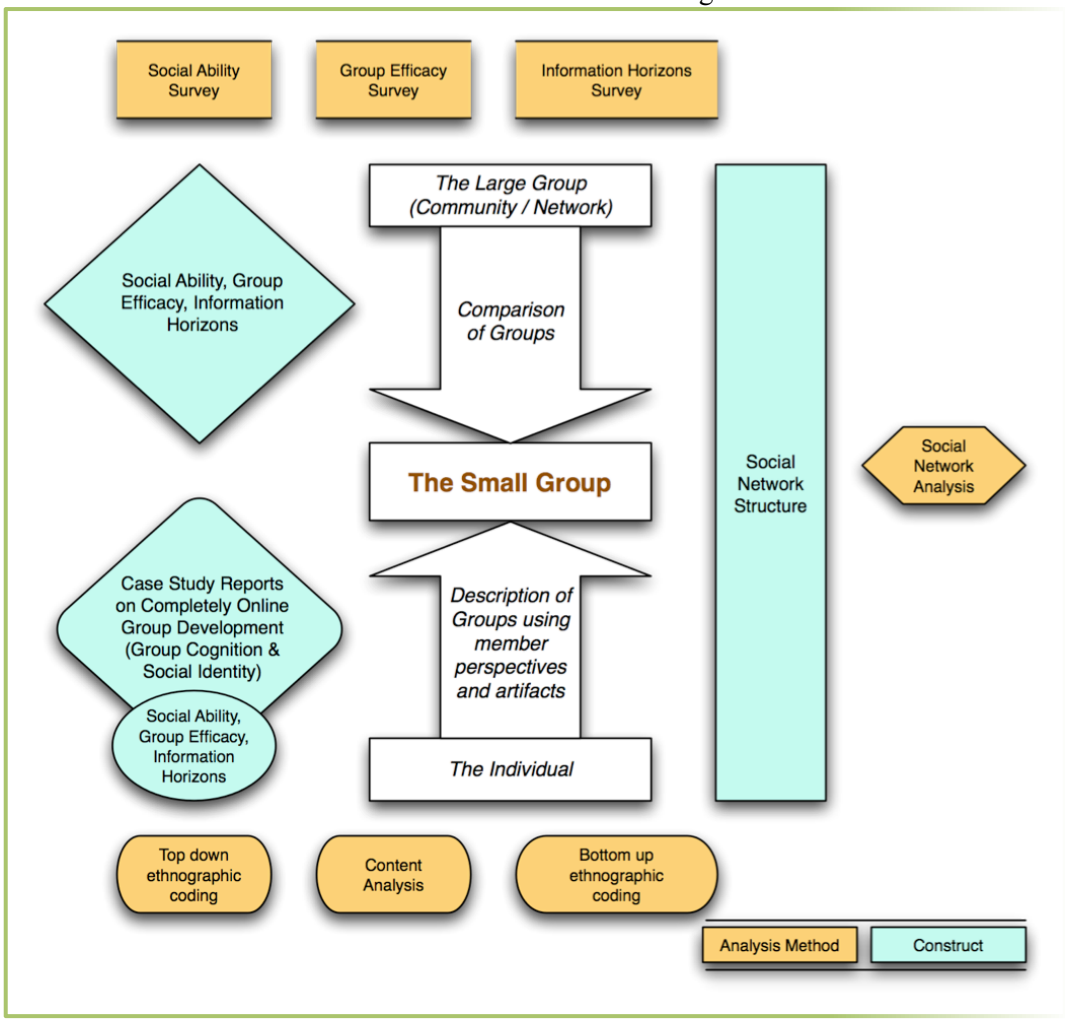


Figure 1 - Data Richness is Sample Data Set

The trace data used in this analysis is possible because of the CANS system.

Trace Data Resolution Levels

This data is trace data from the far right hand side of the diagram above. It is the network trace data we derive from the CANS system. We will walk through the data moving from its lowest resolution toward its highest resolution. In this case, by resolution we mean metadata richness. Table one enumerates this progression.

Cite as: Goggins, S. P., Mascaro, C., & Valetto, G. (2013). Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology*, *64*(3), 516-539.

Table 1 - Data Resolution Levels for CANS Data Analysis

Data Resolution Level	Data Resolution Level Description
(1) Raw CANS Data	One event per row. We know what object (Discussion Board, primarily) the event is pointing toward, and what user created the event. We also know whether the event is a read or post event.
(2) Bi-Directional CANS Data	Most of the analysis reported by Goggins and his co-authors focus on this data set, because it indicates who is reading and posting in response to whom else. For example, if I read a discussion board topic that you created, then a connection is drawn between you and I. In addition to the data in "Raw CANS Data", this data set contains: 1. Weight_in_minutes, which shows the distance in minutes between an event and the object (usually a discussion board) that event is in response to. 2. Object_creator, which shows an anonymized identifier for the object creator.
(3) Exploded Bi-Directional CANS Data	Exploded Bi-Directional CANS data, as explained in Goggins, Laffey, Amelung & Gallager (2010b) and referenced in Goggins, Galyen and Laffey (2010a) recognizes the social form of online discussion to include recognition that when an individual participates in a discussion board in an online course, they attend to more recent posts in that discussion board. This data set adds rows that account for these to the data in the DW_Event_Fact table in the Interaction Warehouse. These rows are distinguished from the non-exploded data using the field row_type: • 0 = an original CANS event row • 1 = an exploded CANS event row
(4) Weighted Extraction of Exploded Bi-Directional CANS Data	This level incorporates findings from grounded theory coding of interviews that are part of Sean P. Goggins dissertation data corpus. Goggins found a consistent two day window of timestamp significance between events in CANS. Following a two day window, the likelihood that a post will be read or that the response is related to the discussion post that is more than two days in the past drops off significantly. This finding can now be applied recursively to the log data – something Goggins did not do as part of his dissertation work – to extract the most significant interactions (those within two days of each other) with a much greater weight than the least significant events.

Level 1 – Raw CANS Data

Raw CANS data shows us basic descriptive pictures, such as the overall degree of participation among members of an online course. We can see in figures xxxSG and xxxSG that one group, called "Get Along" group has the highest overall level of participation, in terms of total event count. Altogether, this group generated over 4,000 CANS events, with a single member of the group (Tommy), generating over 2000 as an individual. Looking at figure xxxSG and xxxSG together, it becomes clear that the group's outlier status originates with Tommy. We can also notice here that Individualist Group member Justin is

Cite as: Goggins, S. P., Mascaro, C., & Valetto, G. (2013). Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology*, *64*(3), 516-539.

substantially lower in his overall participation; compared with everyone else.

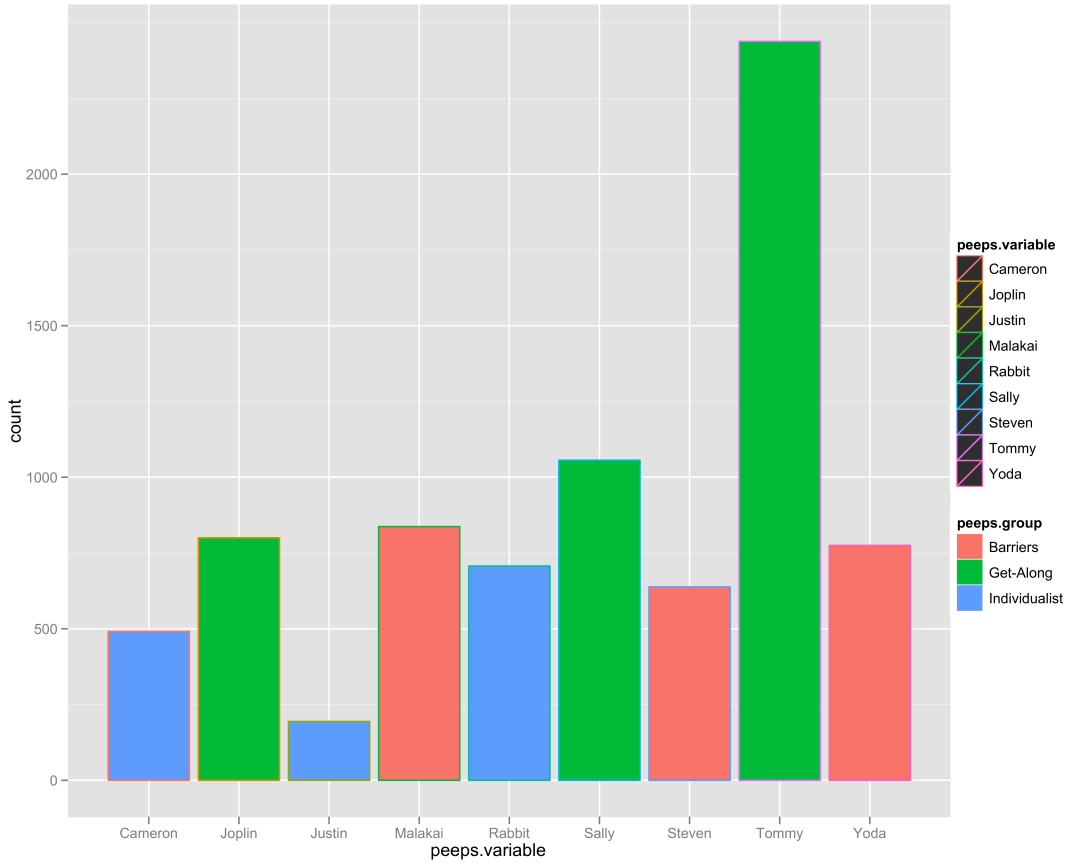


Figure 2 - Total Participation by each member of three case study groups.

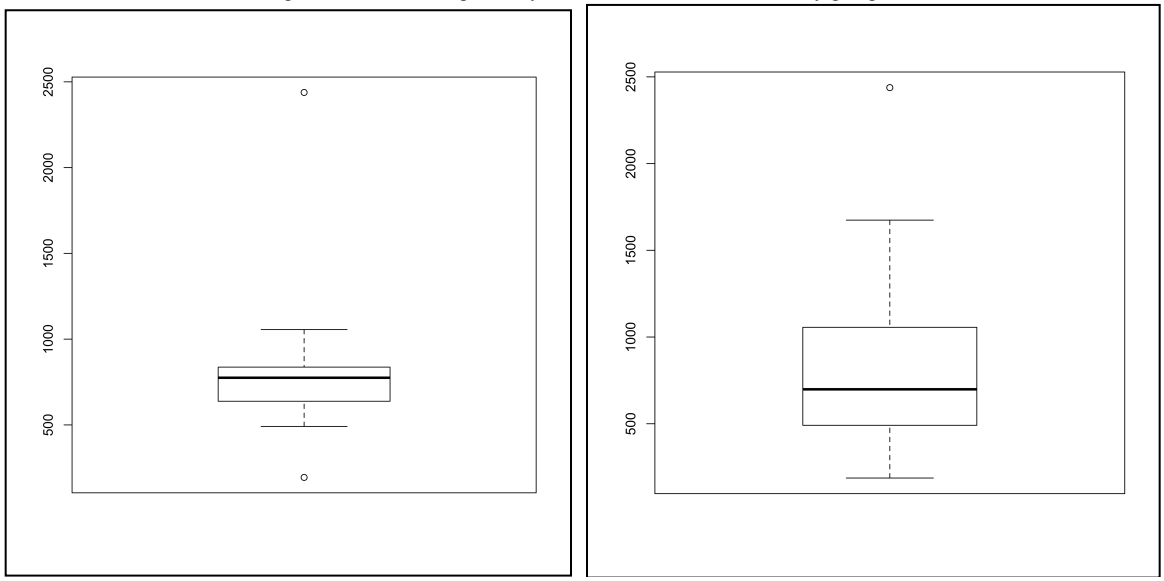


Figure 3 - Boxplot of total event range in the three case study groups on the left, and for the whole course on the right. You can see that the case study groups fall into a narrower range of participation, on average, but the mean participation, symbolized by the black line in both plots, is similar. The case study groups are a good cross section of the course; though theoretically sampled for maximum contrast.

Cite as: Goggins, S. P., Mascaro, C., & Valetto, G. (2013). Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology*, *64*(3), 516-539.

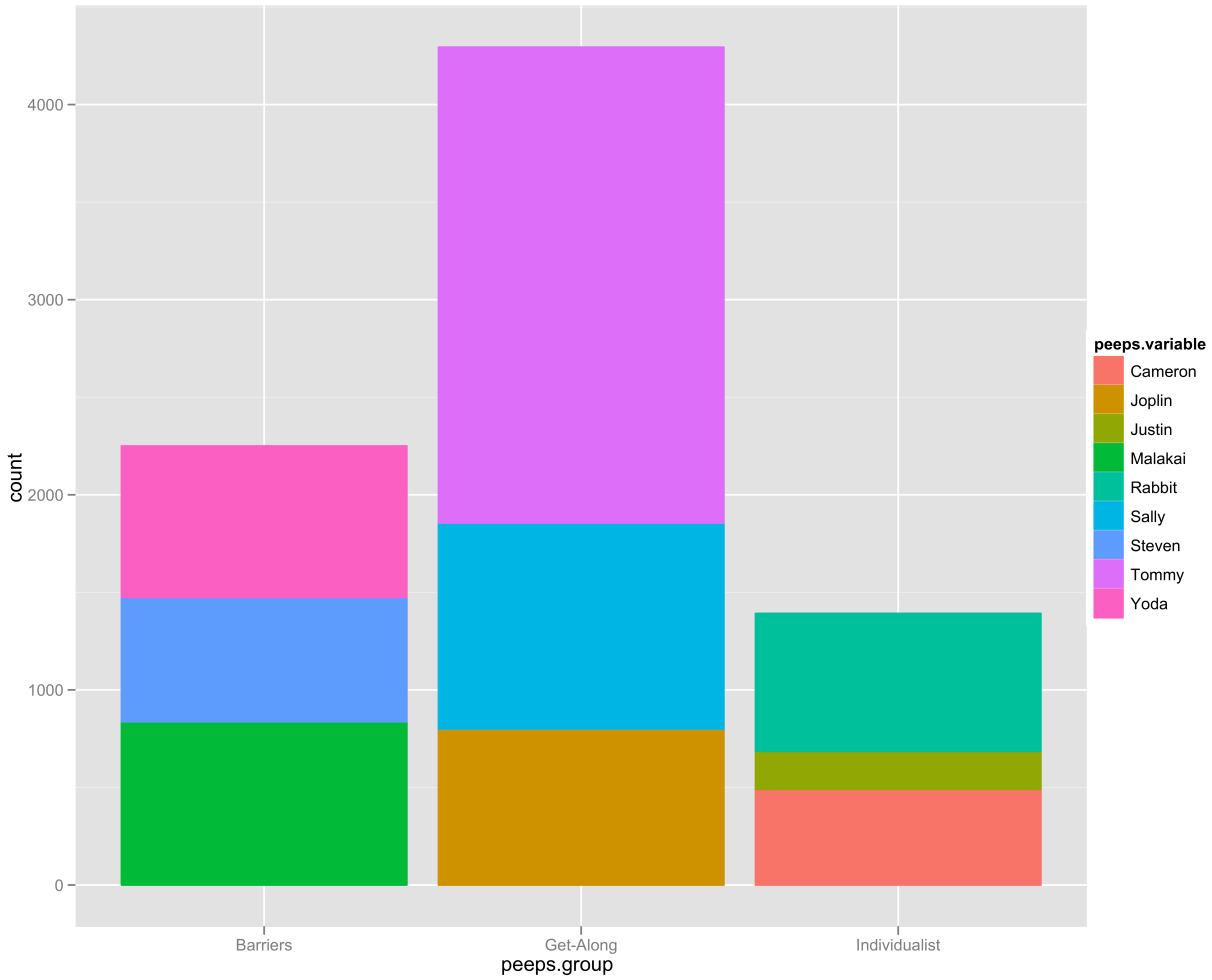


Figure 4 - Total Participation by each member of three case study groups, stacked by group to compare total group activity.

Level 2 – Bi-Directional CANS Data

With Bi-Directional CANS Data, there are a few things we can start to see right away by graphing weight_in_minutes. A cycle becomes apparent from these graphs of the raw, bi-directional CANS data. There is a slow and steady rise in the number of minutes between posts in the bi-directional data. This is because the basic measure of minutes for a discussion board is calculated using the formula: First Topic Post Timestamp (FTPT), and calculates the distance between those timestamps and the timestamp of the event. Often, there is a new discussion board for each module. So, the serial climbing of time distance between events and the posts they are originally responding to is logical. This is also the reason for exploding the bidirectional data. This effect is shown in figures xxxSG and xxxSG.

Cite as: Goggins, S. P., Mascaro, C., & Valetto, G. (2013). Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology*, *64*(3), 516-539.

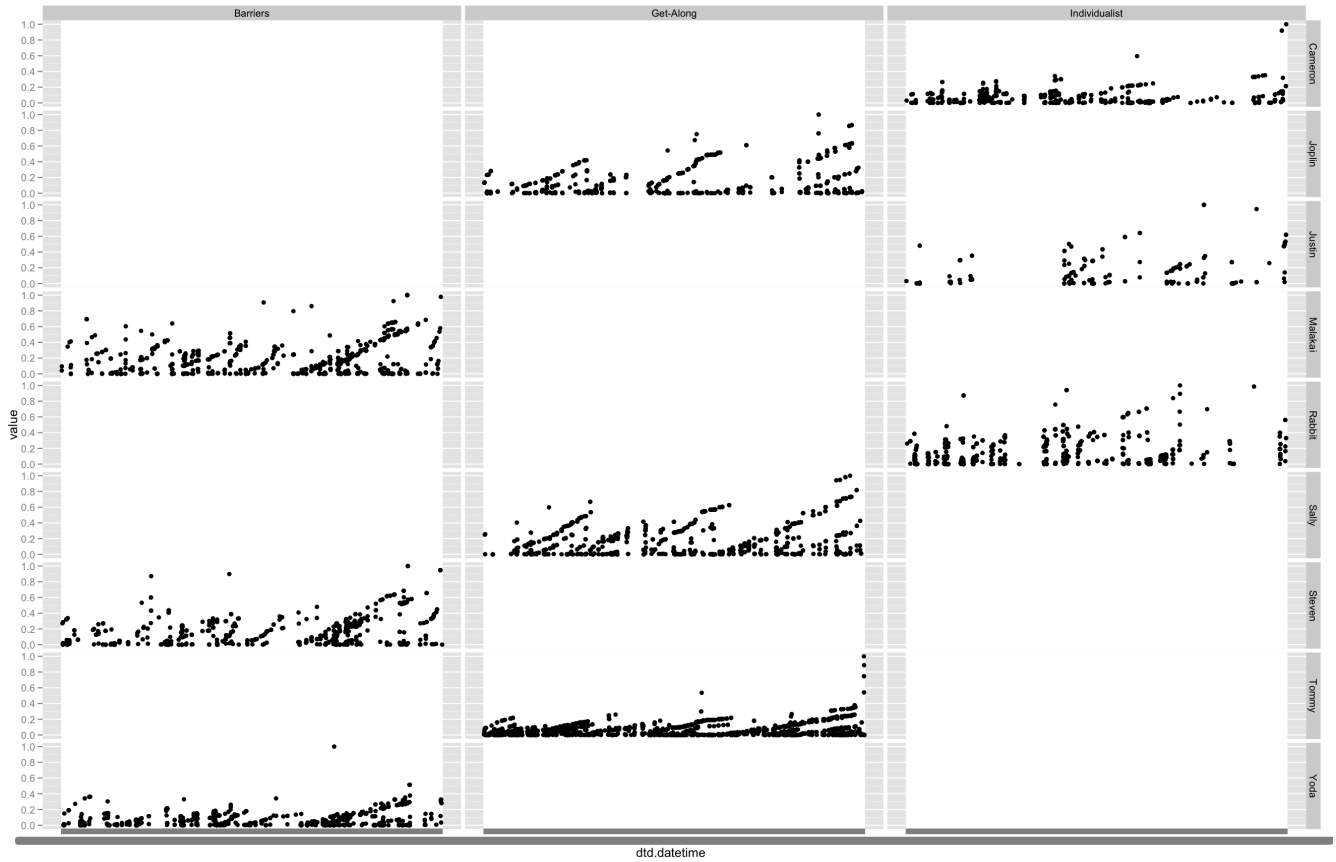


Figure 5 – Weight (time distance) in Minutes for Three Case Study Groups Graphed Over Entire Course

The multi-panel view of the data shown in figure xxxSG reveals that the interaction pattern for Individualist Group is different than for Barriers or Get-Along Group; whose patterns of interaction are similar. This is less clear in the single panel view of figure xxxSG. The cycles for Individualist Group show that, instead of a steady climb during each module like we see for Barriers and Get-Along group, we see fairly fractured collaboration; there are many long breaks in the data for Justin and Rabbit. These breaks are consistent with other data we gathered, including interviews, field notes and surveys. The examination of this trace data, using this type of visualization, provides a clue about the differences in group development observed between the two groups.

Cite as: Goggins, S. P., Mascaro, C., & Valetto, G. (2013). Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology*, *64*(3), 516-539.

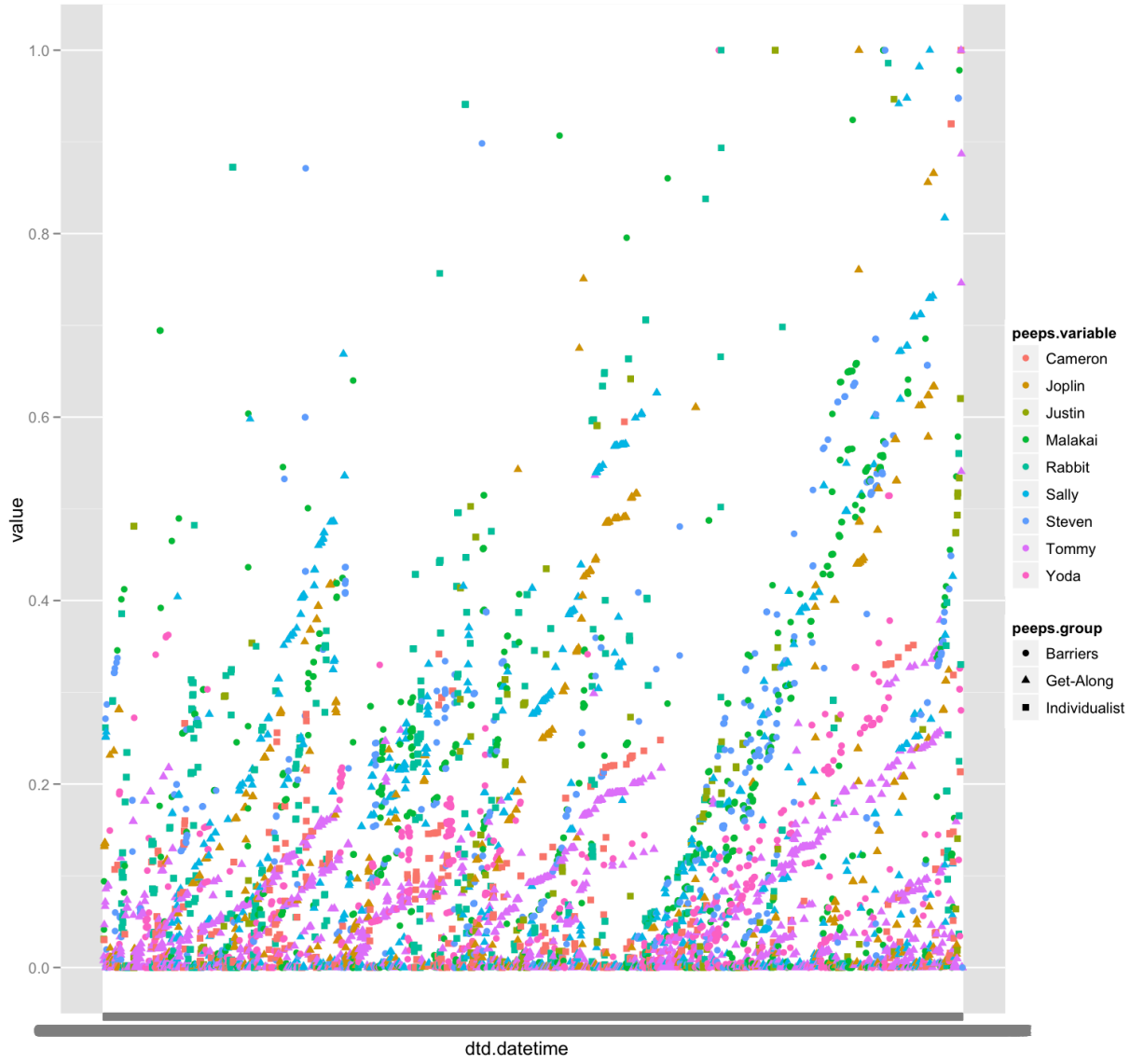
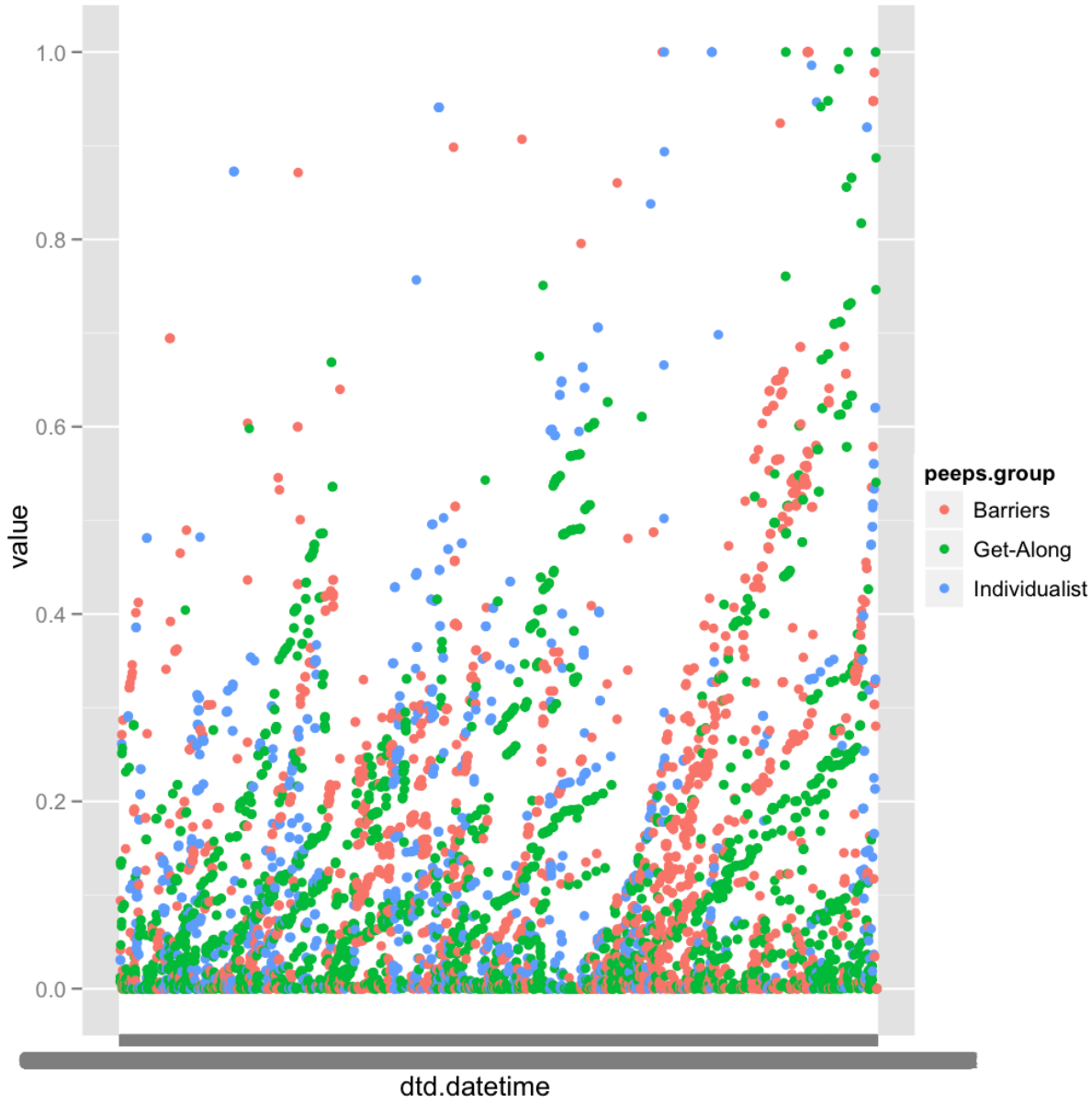


Figure 6 - Weight in Minutes for Three Case Study Groups Graphed Over Entire Course (Single Pane View)

The different pattern of Individualist Group, which is visible in Figure five xxxSG, and less visible in figure six xxxSG. In figure seven xxxSG, the group trend is emphasized, and individual member behavior is not described. Figure seven xxxSG makes the irregularity of Individualist Group member participation visible, but less easily so than figure five. We present the additional figure here to compare and contrast the types of visualizations that might be useful when working with CANS data in different contexts.

Cite as: Goggins, S. P., Mascaro, C., & Valetto, G. (2013). Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology*, *64*(3), 516-539.

`



[1]Figure 7 - Level 2 Group Trends and Cycles

Level 3 – Exploded Bi-Directional CANS Data

The Why and How of Exploding the Data?

When we "Explode" the bi-directional data, we create one row with a time weighted measure (in minutes) between each post in the discussion board, and each post that came before it. The structure of the discussion boards drives this calculation. In other systems, the human-human, human-information and human-computer interaction might be structured differently. The cyclical view we seen in figures

[1] The initial explosion creates a "Time distance weight", which means that the visualizations in this section and the one prior are inverted, compared with Level 4 network graphs, described later.

Cite as: Goggins, S. P., Mascaro, C., & Valetto, G. (2013). Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology*, *64*(3), 516-539.

xxxSG and xxxSG may not be present in MyLyn or VMT data, for example. How the explosion works is explained in Goggins, Laffey, Amelung & Gallager (2010b) and referenced in Goggins, Galyen and Laffey (2010a), and is illustrated in Figure xxxSG.

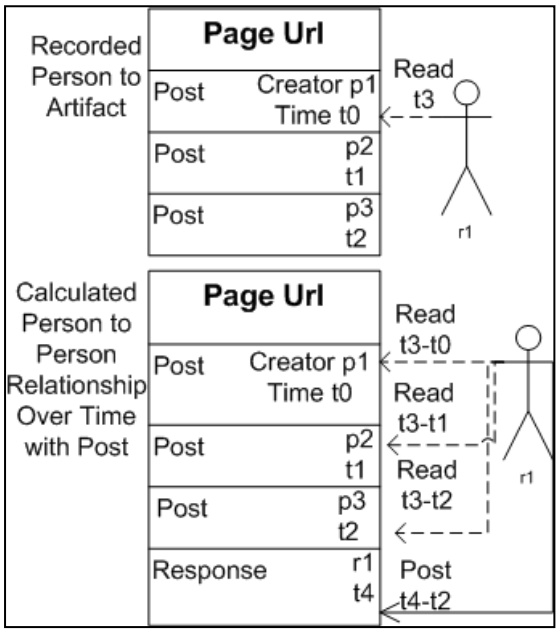


Figure 8 - Mechanics of the explosion process

The data gets "blown up" because in this format we are able to perform more complete social network analysis than we could perform with the Level 2 bi-directional CANS data. Social network analysis was used to address our first research question, which is how and to what extent interactions and group structure vary by activity type. Our interaction analysis was guided by structural theories of social organization [3,6,18], and constructed from CANS data. Individual actions (reads, posts and responses), which occur mostly (92% of events) in the discussion board, are the core data that define network structure in our analysis. The discussion board we use is of the type where up to five posts are viewable on a page. If "reader 1" views a page in a discussion thread, a network "read" tie is created between that person and each person who has previously posted content to that page. Figure one describes this, with p1..pn being a "poster", and t1.. tn being a timestamp. Read actions in other parts of Sakai (8% of events) require a "click" and are recorded as each event occurs.

When we explode the data, we expect the basic shapes of activity count (participation) from Level 2 to remain. We are simply adding implicit connections from the discussion board. For example, Justin has the lowest participation and Tommy the highest participation at both Level 2 and Level 3. Figures ten and eleven xxxSG both demonstrate that the shapes for total participation are uninfluenced by the explosion of the data to meet discussion board needs.

Cite as: Goggins, S. P., Mascaro, C., & Valetto, G. (2013). Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology*, *64*(3), 516-539.

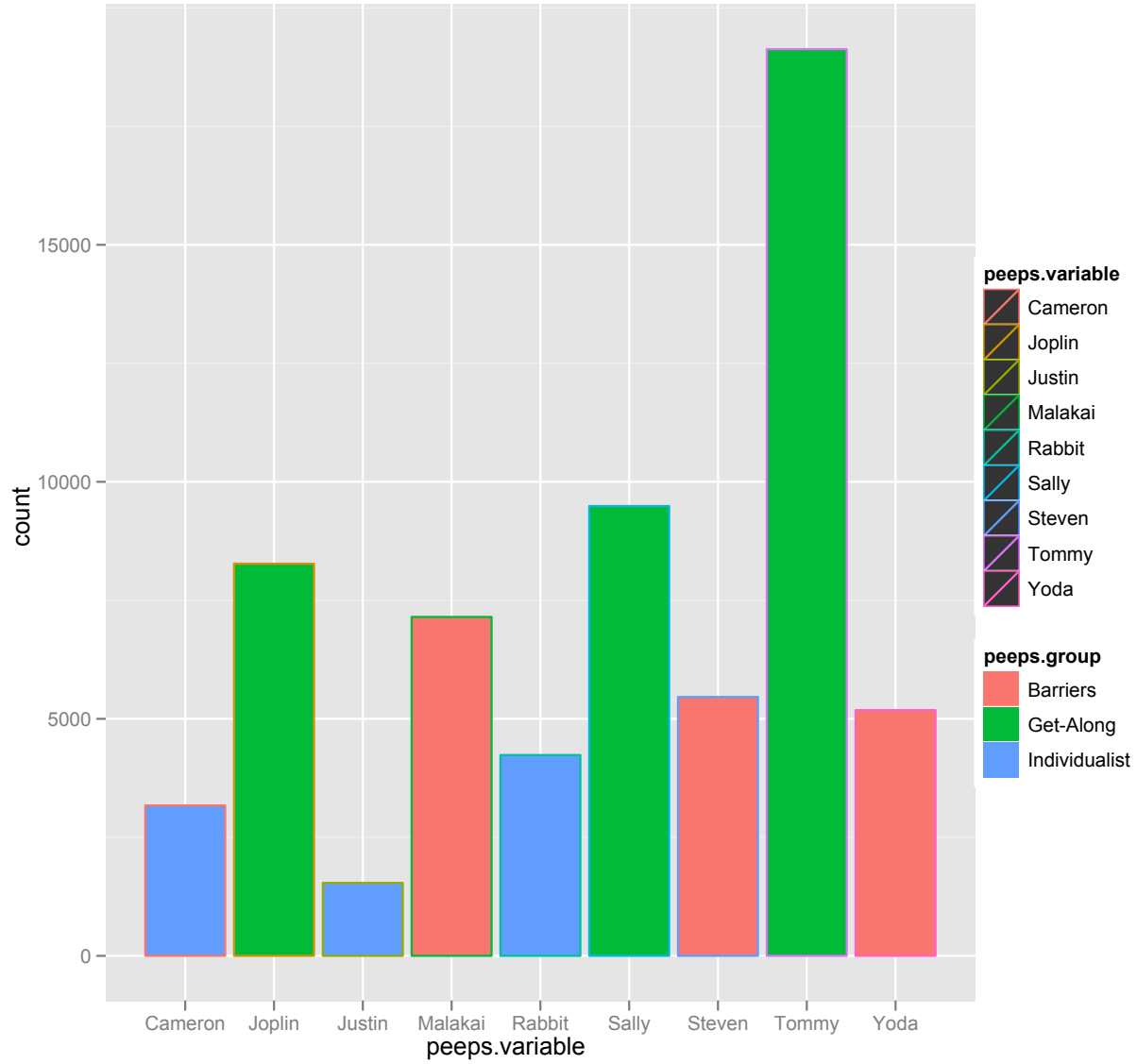


Figure 9 - - Total Participation by each member of three case study groups (with exploded data)

The Data at Level Three does not change at a summary level like we have presented here, but it does create a fuller picture of the interactions that occur, both passively (reading) and actively (posting) between all members of an online course. Without the explosion, out network analysis would be limited by the relationship between an event and the event either immediately preceding it, or the event that initiated the conversation thread. Either choice mutes the true nature of the group structure that develops in a completely online course.

Cite as: Goggins, S. P., Mascaro, C., & Valetto, G. (2013). Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology*, *64*(3), 516-539.

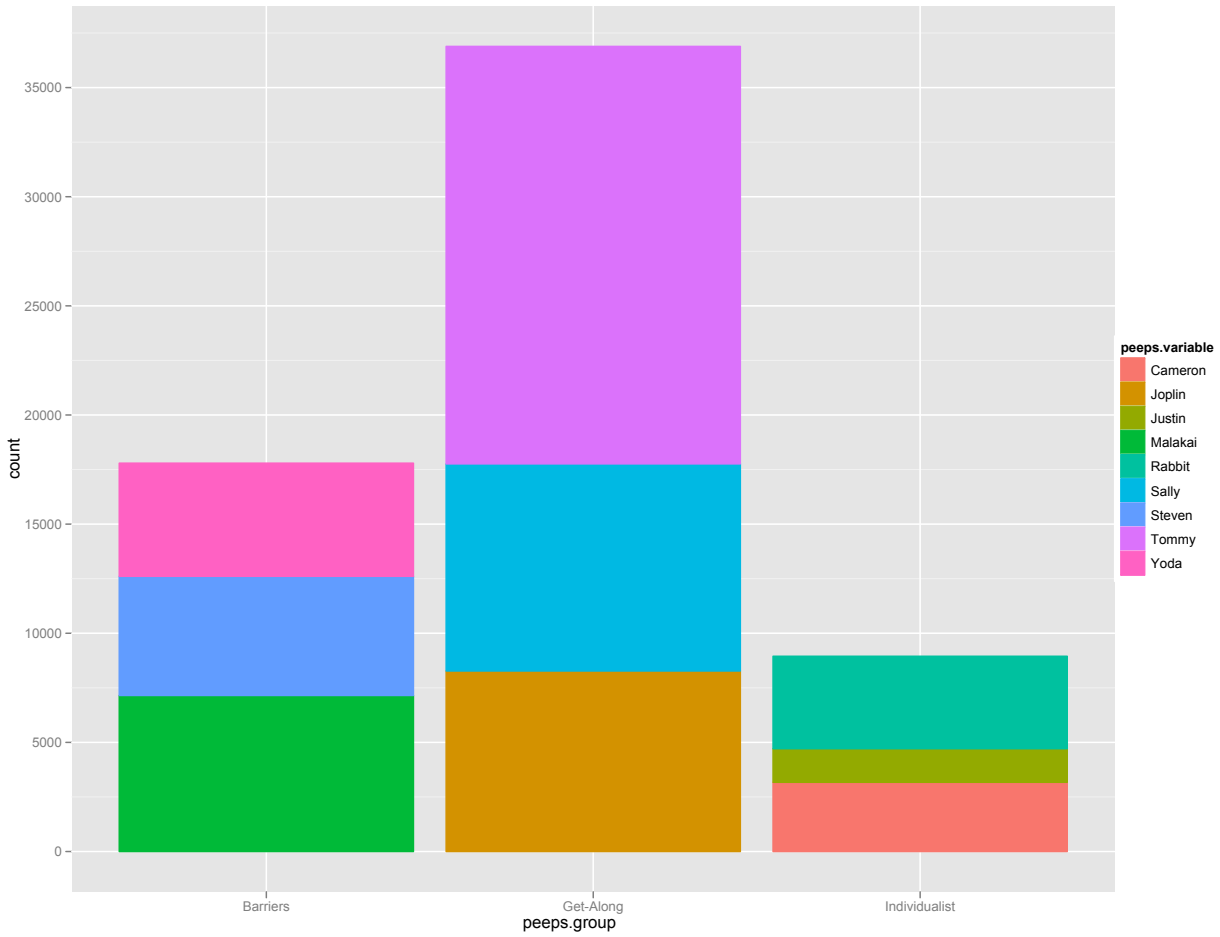


Figure 10 - Total Participation by each member of three case study groups, stacked by groups to show total contrast. (Using Exploded Data).

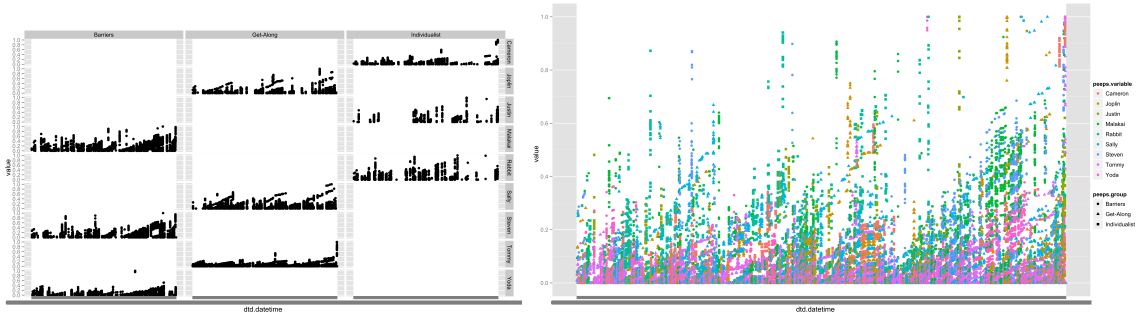


Figure 11 - Basic Shapes of Jitter Plots Remain Consistent from Level 2 to Level 3

Comparison of Network Analysis Using Level Two and Level Three Data

Level 4 – Weighted Extraction of Exploded Bi-Directional CANS Data

There are differences in how the weighted extraction occurs in different source data sets. This is because the weights are semantically different. Response cycles in an online course are predictable and follow a rhythm. Bug response data, in contrast, follows a different rhythm. We will provide examples of how we conceptually and physically extract data from two different projects – the case study of a completely

Cite as: Goggins, S. P., Mascaro, C., & Valetto, G. (2013). Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology*, *64*(3), 516-539.

online course, which we have discussed exclusively to this point; and the MyLyn open source project, which represents a second set of important online data.

Online Course Data

MyLyn Data

When we analyze Mylyn data, time is experienced differently. There are not the firm deadlines, begin times or end times that occur in an online class. In this case, visualizing data over the course of years is important. Here, we look at five different bugs, and when the most conversational activity occurred around those bugs. in order from bottom to top, the bugs are:

1. d
2. d

Weight, however, exists in a conceptually similar way. The calculations of weight for the MyLyn data reflect a time distance that is more linear; with more time passing between communication serving as an indication that the intensity of back and forth conversation is low; and on a per bug, relative scale, the most heavily weighted (shortest time distance) communication is likely to occur as people solve problems related to the bugs. This pattern is shown in figures 12 and 13 (xxxSG).

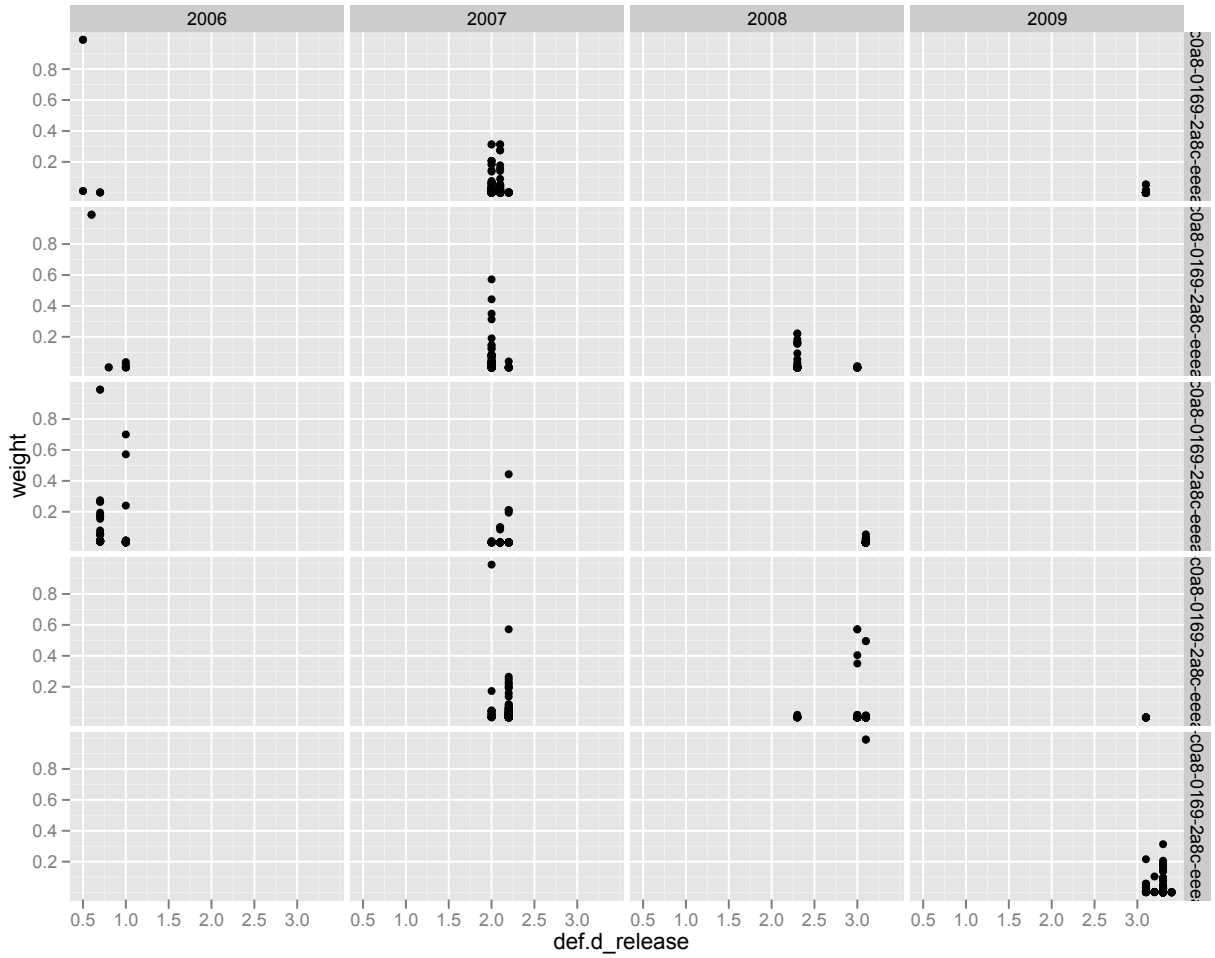


Figure 12 - Five most active bugs by year of activity

Cite as: Goggins, S. P., Mascaro, C., & Valetto, G. (2013). Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology*, *64*(3), 516-539.

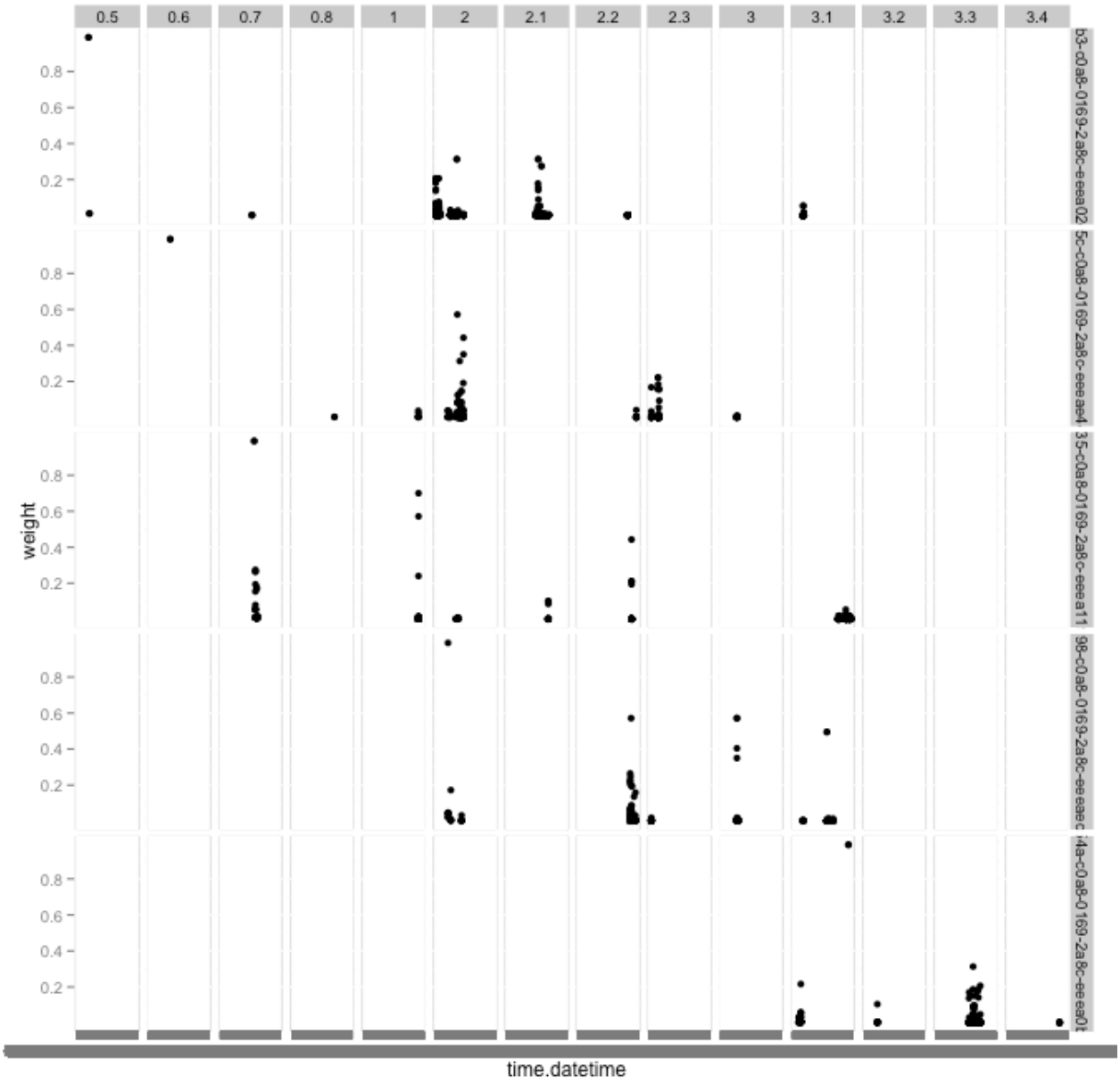


Figure 13 - Incorporating Release and Time

An Example Analysis Process for Mylyn Data

To extract Mylyn Data, there are a number of steps we follow, depending on the specific data. For network analytic data, we have a MySQL stored procedure that will pull data for all of the releases. Simply execute the following command: call cans_warehouse.research_mylyn_release_extract_driver from a database prompt in the CANS_WAREHOUSE database.

References

Goggins, S., Galyen, K., & Laffey, J. (2010a). *Network Analysis of Trace Data for the Support of Group Work: Activity Patterns in a Completely Online Course*. Proceedings from ACM Group 2010, Sanibel Island, FL.

Cite as: Goggins, S. P., Mascaro, C., & Valetto, G. (2013). Group informatics: A methodological approach and ontology for sociotechnical group research. *Journal of the American Society for Information Science and Technology*, *64*(3), 516-539.

Goggins, S. P., Laffey, J., Amelung, C., & Gallagher, M. (2010b). *Social Intelligence In Completely Online Groups*. Proceedings from IEEE International Conference on Social Computing, Minneapolis, MN.