Summary: Data from IBM Jazz project has been analyzed to determine the effects of congruence – and it was found that congruence did not have any effect. Specifically coordination needs were calculated based on who has contributed to which work items, per change sets and communication needs by who has commented on which work items. It was surmised that Jazz developers do not always perform explicit communication (comments), but implicitly communicate via the awareness mechanisms provided by Jazz.

Notes to Authors:

This paper is very well written and analyzes a complex set of data. The discussion of the paper that suggests that all communication need not be explicit and can be facilitated to a great extent through awareness tools rings true. The argument in the paper that congruence measures currently do not take into account communication that is between individuals who do not share coordination needs is an important point. As recognized by Ehrlich et al, in their MSR 07 paper, congruence measures need to identify whether brokers or managers are closing the communication loop.

That said, I have several concerns with the data analysis in the paper. I believe that the analysis of the data is incorrect and the authors need to revise their analysis steps. At this point, I encourage the authors to continue their significant effort, but compare across comparable (complexity, distribution) builds to investigate the effects of congruence.

Detailed comments:

Section 7.1: From what is seen in the figure (Fig 4), it seems that there is indeed some difference in the data – the averages are different – for OK builds, and there is consistently more instances of higher congruence. Although, these differences are seen from the box plots, the Mann-Whitney tests doesn't show that the two samples are statistically different, and that's all that can be said. Extrapolating that result to say that the two samples are equal is wrong. The authors have used the argument that because the two samples cannot be shown to be statistically significant (with 95% confidence), it means that the samples are equal. The test statistics say that the chances for rejecting the null hypothesis (samples are equal) is 34%, which is far lower than 50% of equal chances. Therefore, all that can be said of the Mann-Whitney statistics is that one cannot be absolutely sure that the two samples are from different population (p=0.34), but it doesn't mean that the two samples are actually equal.

Section 7.2: Gaps: It is unclear why extra communication is being penalized. It might be the case that diads have lower congruence, but have a broker that helps with the communication and in fact, the triad has met their coordination needs. To make this information clear, the authors need to show the gap data with and without penalizing (assigning score =-1) the extra communication. Without, this information it is difficult to understand the effects of gap size, because the gap size can be affected in any number of ways.

Section 7.3.1: continuous builds involve collocated teams, which also could mean that team members were communicating informally (meetings, face to face discussions, hall way meetings) and probably knew each other well, so as to have shared mental models, which in turn would require less explicit (formal) communication as evidenced through comments on work items. This represents a significant problem with how communication is tracked in this paper. Comments on work items constitute the sole means of communication. Cataldo et al, in their seminal work had considered f2f communication and chats while calculating the communication behavior. Not considering these channels as well as mailing lists seem to be a major point of weakness in measuring congruence.

I am also curious to see whether OK builds were collocated whereas, Error builds (with more communication) were distributed, which can be the case, since the Error builds involved more work items and larger number of developers. If that is the case, then congruence for collocated teams (OK builds) could be miscalculated, if the collocated teams used other forms of communication.

The authors only mention that the (unweighted) congruence measure was different for the different kinds of builds, but don't provide further details about which builds have higher measure. This information is dealt in passing adding to the cognitive burden on the reader to understand all nuances of the data set.

Section 7.3.2: OK builds seem to be less complicated builds that involved much fewer work items and less coordination needs (page 11, error builds require 2.4 times the coordination needs than OK builds), as compared to Error builds. The paper is comparing two very different kinds of builds here – OK builds that seem to be less complicated, possibly more routine builds with much fewer coordination needs as compared to more complicated builds that needed a lot more coordination. It is not surprising that the more routine, easy builds had less communication, but yet were successful, whereas the more complicated builds had more communication, but were not successful. The entire paper hinges on the comparison of these 2 kinds of builds, which is incorrect because they are comparing builds of very different complexity. To be unbiased, the authors need to choose Ok and Error builds of similar complexity and coordination needs and then identify the effect of congruence on the quality of the build.

I had also a concern with the way congruence is calculated. I would like to know, what is the congruence measure when there is no coordination need (0) and it has not been met (0). The way traditionally congruence was calculated this would cause the value of congruence to be 0 and unmet.

Section 7.3.3 :Error builds have a much higher mean number of work items, which means that the build probably was complex regarding the amount of changes included (could involve complex interdependence among the different work items) as well as the number of people who needed to coordinate each others' work. This goes to show that the error builds probably had a much higher coordination need (as mentioned in Section 7.3.2) and probably despite the increased amount of coordination still had errors, probably because of the complexity involved. Comparing such builds to OK builds, with much fewer work items is incorrect. Figure 9 implies that Error builds seem to have around 3 times the number of work items than OK builds.

Section 7.3.4: I am having difficulty in understanding the point this section tries to make. As mentioned earlier, it seems that Jazz developers do use the commenting functionality. Table 1, does not help in understanding whether a larger percentage of people have commented in OK (137) or Error (60) builds, since there is a big difference in the number of OK and Error builds, this table would have made more sense if they were normalized either based on the total number of build of each kind or by the total number of constituent change sets.

Section 7.3.5: This is a better measure than the earlier proposed measure. This helps compare within subjects (within ok and within builds). Further, this helps reduce the uncertainty of complexity across two builds (# of work items), since the comparison unit is work items and not builds. Given this premise, Table 3 (totals) shows that for Congruence =1: the probability of getting an error build (at the work item level) is 24% (366/(366+1125)) and getting an OK build is 75% (1125/(366+1125)); when congruence =0: the probability of getting an error build is 60% and getting an OK build is 40% → this shows that when there is high congruence there is a higher probability of getting ok builds (75%) as compared to when congruence is 0 (40%). While I cannot say anything about statistical significance of this data, this goes against the main argument of the paper and seems to suggest that congruence actually had an impact on the OK builds. The same trends are seen when we look at the results for No comments and Comments in Table 3.

Finally, the communication behavior as treated by the paper only captures comments on work items, not other forms of communication and not by third party individuals. This hints that the current congruence metric does not capture all communication in the team – as shown by Table 3 – comments left by individuals who do not have technical dependencies commenting (brokers, managers, experts) .

Minor comments: References are not alphabetically sorted.