I had earlier noted that Dr. Chen’s papers on making time series clustering meaningful have some serious problems (the full text of my earlier remarks are below). Recently additional problems have come to light. While Jason admits the problem, I strongly feel that he has not done enough to inform the community the problems, and this could cause others to waste time pursuing fruitless leads. This missive is an attempt to redress this.

The central experiment in Dr. Chen’s paper was allegedly on the famous Cylinder-Bell-Funnel (CBF) data. The formula for creating this dataset appears in over 20 papers, the UCR archive has given the code to create this data more than 1,200 people over the last 5 years. However, Dr. Chen chose to create his own code for creating this dataset. I noted that in the conference version of his work his CBF dataset clearly does not have the right amount of noise (see below). I pointed this out to him on the day the conference version of the paper was published, yet he has never explained this, and when he later published the journal version of the paper, he used the same “noise-lite” version of the data without explanation or comment.

At the end of 2006, Jason emailed me to ask for my data, I asked for his data in reciprocation and he obligingly sent me his Matlab code to create the “CBF” data. I was stunned to see that his version of “CBF” does not have any variability in the time axis, a critical factor in the original version, and indeed the entire reason for its creation. When I pointed this out to Dr. Chen, he said “I didn't pick that up from your paper”, however my paper clearly states “While each time series is of length 128, the onset and duration of the shape is subject to random variability”. It is difficult to explain how dramatically easier it is to deal with a non-varying CBF, so below I pasted a figure to show the difference it makes in the whole clustering case.

As you can see, in the original version on the left, the random variability is enough to confuse simple hierarchical clustering. In contrast, the crippled version of “CBF” is trivial to cluster, even with the full complement of noise (Dr. Chen uses less noise for some unexplained reason). Note how well defined the three clusters are in this case.

In summary Dr. Chen’s papers claims to be able to cluster the CBF dataset, but he was actually clustering a MUCH easier dataset with less noise and no “warping”. Since this experiment is the cornerstone of his papers, this is a huge problem. I suggested to Dr. Chen that he needed to let the community know that the claims made in his paper are not true, he did add the line “Here we construct a version of this time series where the features have fixed onset and duration.” However this line is casually added to the Introduction section, and the rest of the paper still talks about the CBF dataset, for example Figure 10 is labeled “Final, correct clustering of the BellCylinderFunnel data series”. This is VERY misleading. Dr. Chen is not clustering the BellCylinderFunnel dataset, and the casual mention 6 pages earlier that line “Here we construct a version of this time series where the features have fixed onset and duration.” does not absolve him of misleading his readers.

To summarize: Dr. Chen’s papers claim that “that sequential time series clustering can indeed be meaningful”. However the experiments that support this claim are not done on the simple toy example everyone used to discuss this problem, but a much simpler problem crippled to have less noise and no variability. So the correct claim for the work should be: If you have the most trivial problem possible, and you know the value of K, and you know the length of the patterns, then Dr. Chen’s subjective evaluation of his irreproducible experiments are that his “that sequential time series clustering can indeed be meaningful”. This I completely agree with.

Eamonn Keogh.
A rebuttal of “Making Subsequence Time Series Clustering Meaningful”

I believe that a paper in this year’s ICDM misrepresents my work in ICDM 2003. Since conferences don’t offer the possibility to write a “letter to the editor” I am presenting a rebuttal here. Note that this is not a personal attack; I just want my work represented fairly.

Jason’s text: “…sequential time series clustering can indeed be meaningful, and that it is not, as recommended by Keogh et. al., intrinsically flawed by definition.”

Eamonn’s rebuttal: I never “recommended” (or suggested, or stated) that sequential time series clustering cannot be meaningful. In my paper I stated, “We have shown that a popular technique for data mining does not produce meaningful results”. My work spoke about a particular popular technique. In the same paper I introduced a tentative technique for sequential time series clustering. If I introduced a method to solve a problem, is it likely I believe that it cannot be solved? In addition, I successfully demonstrated my technique on the same problem as Jason, except my data had the correct amount of noise, and Jason’s version mysteriously has much less noise (see below).

Jason’s text: “We show that, using this alternative distance measure, sequential time series clustering can indeed be meaningful”

Eamonn’s rebuttal: In fairness, Jason shows one irreproducible experiment¹, on a toy problem. Furthermore, although he uses the well-known CBF dataset, Jason’s version of the data seems to have much less noise, (see the figures below) for reasons that are not explained. Finally, the result is for only a cyclic dataset, which are rare and can be trivially handled with Fourier techniques etc.

¹ Jason wrote to 3rd party working on this problem, who wanted to reproduce Jason’s results “I don’t have the data set in a file as such, I reconstructed it on the fly for each experiment.”
Finally, I am open to the possibility that any of my work could be flawed, and I would be delighted to correct such mistakes. But the work in question is simply misrepresenting my words without any challenge to the technical results. However I am not suggesting any malice by the author, and wish him the best of luck with this research.

Eamonn Keogh.

General note: For some reason my paper seems to have attracted more that its fair share of misunderstanding and controversy (one of the original reviewers wrote “you will get into bad trouble if you publish this”). Below is an account of some of the dozens of emails I have gotten from people that have empirically verified my claims.

Dr. Loris Nanni noted that she had encountered problems clustering economic times series. After reading an early draft of our paper she wrote “At first we didn't understand what the problem was, but after reading your paper this fact we experimentally confirmed that (STS) clustering is meaningless!!” (Nanni, 2003). Dr. Richard J. Povinelli and his student Regis DiGiacomo experimentally confirmed that STS clustering produces sine wave clusters, regardless of the dataset used or the setting of any parameters (Povinelli, 2003). Dr. Miho Ohsaki re-examined work she and her group had previously published and confirmed that the results are indeed meaningless in the sense described in this work (Ohsaki et al., 2002). She has subsequently been able to redefine the clustering subroutine in her work to allow more meaningful pattern discovery (Ohsaki et al., 2003). Dr Frank Höppner noted that he had observed a year earlier than us that “…when using few clusters the resulting prototypes appear very much like dilated and translated trigonometric functions…” (Hoppner, 2002); however, he did not attach any significance to this. Dr. Eric Perlman wrote to tell us that he had begun to scaling up a project of astronomical time series data mining (Perlman and Java, 2003); however, he abandoned it after noting that the results were consistent with being meaningless the sense described in this work. Dr. Anne Denton noted, “I’ve experimented myself, (and) the central message of your paper – that subsequence clustering is meaningless – is very right,” and “it’s amazing how similar the cluster centers for widely distinct series look!” (Denton, 2003)


